Avertissement
Ébauche – Pour fin de discussion seulement; ne représente pas les positions du G7 ou de ses membres. Ce document a été rédigé pour guider les discussions durant les discussions en petit groupes de la Conférence multipartite du G7 sur l'intelligence artificielle du 6 décembre 2018 à Montréal au Canada.

Document de discussion pour les Séances en petits groupes

Thème 3: Responsabilité en matière d'IA

Promouvoir une plus grande confiance de la société

Auteurs :
Jason Millar, Ph. D., Université d'Ottawa
Brent Barron, CIFAR
Koichi Hori, Ph. D., Université de Tokyo
Rebecca Finlay, CIFAR
Kentaro Kotsuki, IICP
Ian Kerr, Ph. D., Université d'Ottawa,
président du conseil Intelligence artificielle et société du CIFAR

Conférence multipartite du G7 sur l'intelligence artificielle
6 décembre 2018, Montréal

Ébauche – Pour fin de discussion seulement; ne représente pas les positions du G7 ou de ses membres.

Sommaire

Le présent document a été élaboré à la demande du gouvernement du Canada pour appuyer la Conférence multipartite du G7 sur l'intelligence artificielle : permettre l'adoption responsable de l'IA le 6 décembre 2018. Des coresponsables du Canada et du Japon ont élaboré ce document sur la responsabilité, dont l'objectif est de fournir un point de départ pour les discussions sur le sujet de la *Responsabilité en matière d'IA : promouvoir une plus grande confiance de la société* dans le cadre de la conférence. Le présent document et la discussion s'appuient sur les travaux entamés lors de la rencontre de 2016 des ministres des TIC à Takamatsu et qui ont mené, plus récemment, à la Vision commune de Charlevoix sur l'avenir de l'intelligence artificielle[1].

Le présent document est divisé en deux sections. La première fournit des renseignements sur les travaux réalisés à ce jour dans le domaine et présente divers concepts et distinctions qui méritent d'être soulignés lorsqu'on pense à la responsabilité et à la confiance dans l'IA. La deuxième section traite du processus de consultation et des mesures qui pourraient être prises par différents groupes d'intervenants pour l'avenir.

Remerciements et notes des auteurs

Nous remercions sincèrement le CIFAR et le ministère des Affaires intérieures et des Communications du Japon, ainsi que les nombreux intervenants du Canada et du Japon qui ont fourni des commentaires et des conseils sur ce document. Veuillez noter que les idées exprimées dans le présent document sont celles des auteurs ou ont été fournies dans le cadre de la consultation des intervenants. Ce ne sont pas celles des gouvernements du Canada ou du Japon.

CIFAR

情報通信政策研究所
Institute for Information and Communications Policy

Questions à discuter lors de la rencontre du G7

Sept questions, regroupées sous trois grandes catégories, sont proposées pour encadrer les discussions de la conférence :

Principes

- Q1 : Quels sont certains principes communs pour la responsabilisation de l'intelligence artificielle (IA) dans tous les secteurs?
- Q2 : Comment pouvons-nous déterminer quels systèmes d'IA nécessitent des régimes de responsabilisation plus rigoureux pour une gouvernance appropriée?

Développement

- Q3 : Étant donné que la confiance peut être mal placée – les personnes peuvent avoir trop ou pas assez confiance en l'intelligence artificielle –, comment les régimes de responsabilisation peuvent-ils favoriser le développement d'une *IA fiable et digne de confiance*?
- Q4 : Comment établir un équilibre entre la responsabilité et l'innovation afin que les avantages de l'IA soient garantis de manière responsable et inclusive?

Instruments

- Q5 :Comment pouvons-nous assurer une pluralité représentative et diversifiée de voix et de points de vue dans l'élaboration de régimes de responsabilisation internationaux et nationaux pour l'IA?
- Q6 : Quels mécanismes (réglementaires ou non) sont les plus appropriés pour régir les diverses applications de la prise de décisions en fonction d'algorithmes?
- Q7 : Quels rôles les différents intervenants (p. ex. les gouvernements, les organisations internationales, les développeurs privés, les fournisseurs de services et les utilisateurs, le système juridique, etc.) devraient-ils jouer pour assurer la responsabilisation en matière d'IA et la coordination au-delà des limites des territoires et des frontières culturelles?

Travaux accomplis à ce jour

Introduction

Avec le développement et la prolifération des systèmes d'IA, il est urgent de s'attaquer aux questions de responsabilisation. Cependant, il y a un « manque de consensus au sein de la communauté au sens large quant à ce à quoi ressemblerait une "trousse de solutions" »[2]. Le présent document examine la question de la responsabilité en matière d'IA et de son lien avec la confiance, propose des définitions et des distinctions clés et présente quelques considérations pour les discussions futures et les mesures possibles pour les membres du G7, les autres pays et les intervenants du monde entier.

Terme clé : intelligence artificielle (IA)

« [L'IA] consiste à fabriquer des ordinateurs qui peuvent nous aider et faire ce que les humains peuvent faire, mais pas nos ordinateurs actuels. » – Yoshua Bengio [traduction]

« Le secteur de l'informatique dédié à la résolution des problèmes cognitifs couramment associés à l'intelligence humaine, comme l'apprentissage, la résolution de problèmes et la reconnaissance de formes. » – Amazon [traduction]

« C'est la science et l'ingénierie de la fabrication de machines intelligentes, en particulier de programmes informatiques intelligents. Elle est liée à la tâche similaire d'utiliser les ordinateurs pour comprendre l'intelligence humaine, mais l'IA n'a pas à se limiter à des méthodes biologiquement observables. » – John McCarthy [traduction]

Le terme *intelligence artificielle* (IA) englobe un large éventail de technologies et d'approches. Deux approches générales de l'IA méritent d'être distinguées. Une approche utilise des modèles prédéfinis pour atteindre des objectifs, et l'autre repose sur *l'apprentissage machine* pour entraîner un système à atteindre des objectifs. Il existe deux techniques bien connues en apprentissage machine. Pour les définir de manière très générale, il s'agit de *l'apprentissage profond*, qui utilise de très grands réseaux neuronaux artificiels et de *l'apprentissage par renforcement*, qui utilise une structure de récompense et de punition. Le présent document a pour but de discuter de la responsabilité dans son application générale à l'IA, tout en reconnaissant que certaines questions éthiques qui sont devenues associées à l'intelligence artificielle, en particulier l'explicabilité, sont plus directement liées à l'apprentissage profond.

La recherche sur l'IA a progressé rapidement au cours de la dernière décennie. Le succès en laboratoire a conduit à la prolifération de systèmes d'IA dans certains secteurs de la société. En raison de sa capacité à opérer sur des ensembles de données massifs avec une rapidité, une précision et une exactitude qui surpassent les capacités humaines, l'IA commence à être appliquée, ou est envisagée, dans les domaines de la santé, des transports, de l'ordre public, de la défense et des finances – pratiquement tous les secteurs de l'économie – pour appuyer et, dans certains cas, remplacer l'analyse humaine et la prise de décisions. Ces capacités permettent à l'IA d'apporter de grands avantages à la société.

Comme pour toute nouvelle technologie, nous apprenons que le déploiement de l'IA au-delà des laboratoires peut constituer une source de risques pour les personnes et les sociétés, ce qui soulève des préoccupations quant à la responsabilisation. Voici quelques exemples qui aident à illustrer la situation. L'IA qui est formée sur des ensembles de données biaisées peut intégrer et faire proliférer ces biais dans ses extrants, ce qui mène à des applications discriminatoires[3]. Dans la pratique, de nombreux systèmes d'apprentissage profond fonctionnent en grande partie comme des « boîtes noires », de sorte que leur comportement peut être difficile à interpréter et à expliquer, ce qui soulève des préoccupations sur leur explicabilité, la transparence et sur le contrôle humain[4]. La responsabilisation est d'autant plus compliquée par le fait que les systèmes d'IA déployés peuvent être composés de nombreux éléments (code, capteurs, jeux de données, etc.) qui peuvent chacun représenter des points de défaillance potentiels. Enfin, puisque les humains perçoivent souvent l'IA comme étant « supérieure » dans ses capacités, ils peuvent *trop* lui faire confiance, ce qui peut entraîner des blessures physiques et mentales[5]. Ces exemples ne sont pas exhaustifs. À mesure que nous en apprenons davantage sur l'IA et ses caractéristiques uniques, la liste des méfaits potentiels s'allonge. La compréhension de ces méfaits potentiels commence à être intégrée dans la réflexion des gouvernements sur l'IA[6]. En effet, la recherche systématique sur les implications éthiques de l'IA progresse constamment, tant à l'intérieur qu'à l'extérieur du milieu universitaire. Certaines de ces nouvelles recherches visent précisément à aider les décideurs et les ingénieurs à prévoir et à aborder les questions éthiques liées à l'IA, dont la responsabilité[7].

Il est urgent d'anticiper et de traiter ces risques potentiels. Ces systèmes sont souvent opaques et complexes et leur impact potentiel est vaste. Combinés à leur utilisation potentielle dans des contextes décisionnels critiques et à enjeux élevés (p. ex. raisonnement judiciaire, santé, guerre, opérations financières), ils peuvent avoir des répercussions importantes. Par exemple, une mise à jour de routine d'un logiciel à un algorithme d'acheminement du trafic contrôlant un système de mobilité automatisé et connecté pourrait rapidement retransmettre les risques à des millions de personnes dans le système. Déterminer qui devrait être confronté aux risques les plus importants au sein d'un système de mobilité est une lourde tâche qui a de vastes répercussions[8]. Le processus par lequel nous devons prendre cette décision, ainsi que la responsabilité de cette décision et ses conséquences systémiques, peuvent dépasser les capacités des régimes existants (responsabilité délictuelle, protection des consommateurs, etc.)[9].

Même s'il est certainement possible de causer du tort en déployant l'IA dans certains contextes, nos préoccupations devraient rester modérées. Dans de nombreux cas, les répercussions sociales négatives des systèmes du statu quo (c.-à-d. les systèmes non algorithmiques) ne sont pas interrogées aussi intensément que les systèmes d'IA[10]. Autrement dit, il est important de comprendre les risques posés par l'IA ainsi que les risques posés par le statu quo.

Comme c'est souvent le cas, le rythme de l'innovation technique dépasse celui de l'élaboration de politiques en matière de responsabilité. L'absence de directives claires en matière de responsabilité pourrait miner la confiance des experts et du public, ce qui pourrait limiter les avantages de l'IA. En même temps, il est important de noter que l'objectif ne peut pas être simplement d'accroître les niveaux de confiance dans l'IA.

Nous pouvons faire *trop* ou *pas assez* confiance à un système automatisé. Nous nous méfions lorsque des hypothèses inexactes (p. ex. des craintes ou de la désinformation) au sujet de l'IA nous empêchent de lui faire confiance, ce qui pourrait nous priver des avantages qu'elle pourrait produire. D'un autre côté, nous faisons trop confiance à un système lorsque, par exemple, nous croyons (et avons confiance) à tort qu'un système est capable d'accomplir certaines tâches qu'il n'est pas capable d'accomplir. Les accidents malheureux causés par des véhicules autonomes peuvent être considérés comme des cas de trop grande confiance : dans chaque cas, le conducteur humain croyait à tort que le système automatisé de contrôle de la conduite était capable de fonctionner à un niveau auquel il n'était pas capable. Ainsi, notre objectif pourrait être d'encourager des niveaux *appropriés* de confiance dans l'IA, avec des régimes de responsabilisation qui tiennent compte des nuances entre la confiance excessive et insuffisante[11].

Enfin, avec les progrès de la mise en réseau de l'IA, où les systèmes d'IA sont connectés à d'autres systèmes, par Internet ou d'autres réseaux d'information et de communication, il deviendra plus difficile de déterminer à la fois les causes des problèmes et la responsabilité qui en découle. Afin de favoriser la confiance à l'égard de l'IA, il sera important de s'appuyer sur un ensemble de principes communs qui clarifient les rôles et les responsabilités de chaque intervenant du réseau, y compris des développeurs, des fournisseurs de services et des utilisateurs finaux, dans la recherche, le développement et l'utilisation de l'IA.

Responsabilité, confiance et fiabilité

De manière générale, dans notre société, la responsabilisation est à la base de la confiance. La responsabilisation représente la reconnaissance et l'imputation de la responsabilité ainsi que la reddition de compte pour les actions posées, les décisions prises, les produits proposés et les politiques adoptées. Présentement, trois « sens » de la responsabilité existent dans la littérature sur l'IA, chacun pointant dans une direction différente. Le premier des trois veut que la responsabilité soit une caractéristique même du système d'IA[12]. En ce sens, l'intégration de l'explicabilité aux systèmes d'IA permettrait de régler une partie de la question de la responsabilité. Le deuxième vise l'identification des personnes ou groupes responsables de l'incidence des algorithmes ou d'IA[13]. En ce sens, la responsabilité est en quelque sorte étroitement liée à l'identification des personnes responsables des différentes incidences sur le système sociotechnique. Finalement, et possiblement de manière plus générale, la responsabilité est perçue comme une caractéristique du vaste système sociotechnique qui élabore, produit, déploie et emploie l'IA[14]. Par exemple, AI Now propose une cadre d'évaluation algorithmique de l'incidence, semblable à une évaluation des facteurs relatifs à la vie privée, permettant d'intégrer la responsabilisation au sein d'un plus vaste système sociotechnique qui a recours à l'IA, mais dans lequel uniquement une partie servirait à l'attribution de la responsabilité[15]. De manière semblable, la World Wide Web (WWW) Foundation définit les principes de la responsabilisation algorithmique, notamment l'équité, les caractères explicable et vérifiable, la responsabilité et l'exactitude[16].

Ces trois sens de la responsabilité font l'objet de recherches actives et de projets de développement.

La WWW Foundation évoque une distinction « fondamentale » entre la responsabilisation algorithmique, soit la responsabilité qu'ont les concepteurs d'algorithmes de fournir des preuves de préjudices possibles ou réels, et la justice algorithmique, soit la capacité de réparer les torts causés[17]. La raison pour laquelle il est important de faire cette distinction est que de centrer les efforts sur la réparation des torts causés pour traiter de la responsabilisation pourrait détourner l'attention des concepteurs et ingénieurs d'algorithmes de la chance inouïe qui s'offre à eux d'anticiper les préjudices avant que l'IA ne soient rendue disponible. Tout en tenant compte de ces conseils, il est aussi essentiel de ne pas trop mettre l'accent sur la responsabilité des concepteurs d'algorithmes en matière d'anticipation des préjudices, ce qui pourrait distraire d'une approche plus globale à l'égard de la responsabilisation en IA.

Les observations présentées ci-dessus qui ont trait à la confiance et à la responsabilité tendent à faire une distinction bénéfique entre la *confiance en un système* et la *fiabilité d'un système*[18]. La confiance en un système signifie avoir un niveau de confiance suffisamment élevé, ou tout juste assez élevé, en un système. Par conséquent, la *fiabilité* d'un système peut être définie par la capacité d'un système à effectuer ou réaliser de manière fiable l'objectif pour lequel il a été conçu et comme il est attendu. Par exemple, le niveau de confiance en certaines nouvelles circulant sur les réseaux sociaux est souvent plus élevé qu'il ne le devrait, et ce, en raison de leur inexactitude et leur caractère peu fiable. De la même façon, le niveau de confiance envers les publications scientifiques est souvent moins élevé qu'il ne le devrait compte tenu de leur niveau de fiabilité. Donc, il est possible de dire que le niveau de confiance est en-deçà de celui attendu. En guise de dernier exemple, les gens ont raison d'avoir confiance dans le transport par avion puisque toutes les preuves indiquent que ce mode de transport est fiable. Lorsqu'il est question d'IA, divers facteurs entraînent une méfiance des gens envers la fiabilité de certaines de ces technologies, ce qui pourrait limiter les avantages qu'elles présentent. En élaborant un ensemble strict de règles en matière de responsabilisation des systèmes d'IA, dont de plus vastes systèmes sociotechniques environnants, il est possible de promouvoir une *confiance appropriée envers l'IA auprès des experts et du public*.

On cite souvent la transparence dans les discussions entourant la responsabilisation en matière d'IA, puisque cette notion permet un contrôle plus serré des systèmes d'IA. Cependant, on n'améliore ou n'accroît pas la responsabilisation simplement en augmentant la transparence. La transparence est sans aucun doute une composante de la responsabilisation, mais en l'absence de processus, principes et cadres stricts qui font en sorte que cette notion mène à une plus grande responsabilité, elle est nécessaire, mais insuffisante.

Un autre défi pour l'IA responsable est que l'IA a la capacité de traverser les frontières. Elle a été mise au point et déployée à travers différents territoires, et ce, de façon à ce qu'elle traverse non seulement les frontières internationales, mais aussi culturelles. La distribution et le déplacement des biens numériques sont difficiles à contenir. Cela a pour effet de rendre la notion de confiance plus complexe et peut être démontré, par exemple, lorsqu'une technologie d'IA est mise au point selon un ensemble d'éléments culturels et que cette technologie est exportée dans un contexte culturel « étranger », où

les normes de confiance diffèrent. Cela complique aussi les interventions des différentes autorités puisqu'une technologie d'IA pourrait être conçue d'une façon qui ne respecte pas nécessairement les lois ou normes culturelles locales. Bien que les difficultés liées aux éléments relevant des différentes administrations ne soient pas nouvelles, elles caractérisent bon nombre d'enjeux auxquels nous devons faire face dans cette ère numérique, dont la très importante question de la vie privée. Comme nous avons pu le constater récemment avec le règlement général sur la protection des données (RGPD) de l'Union européenne, les solutions pangouvernementales nécessitent la participation de divers intervenants et profiteraient d'une coordination multilatérale. Cette coordination permettrait non seulement de garantir que l'IA respecte les contraintes légales de diverses compétences, mais qu'elle fonctionne de façon sécuritaire et fiable.

Finalement, il faut poursuivre la recherche de manière à mieux appuyer la prise de décisions en lien avec :

- le biais;
- l'explicabilité en matière d'IA;
- l'éthique des processus d'ingénierie et de conception[19];
- les stratégies efficaces en matière de mobilisation du public et des différents intervenants pour la responsabilisation en IA;
- les différentes options efficaces en matière de politiques;
- les enjeux éthiques en matière d'IA;
- les analyses juridiques;
- les stratégies de surveillance et de vérification efficaces en lien avec l'IA et les techniques pouvant s'appliquer pour différents systèmes et technologies;
- le journalisme informatique[20], etc.

Activités internationales

Différents programmes, politiques, activités et centres internationaux ont été mis sur pied pour traiter de la conception d'une responsabilisation rigoureuse et globale de l'IA. Certaines lacunes en matière de connaissance que ces derniers doivent pallier comprennent notamment la façon :

1. d'assurer la participation du secteur public dans la conception des ensembles de règles appropriées entourant la responsabilisation en IA;
2. de soutenir l'élaboration et le maintien d'un ensemble strict et global de règles entourant la responsabilisation en IA;
3. d'élaborer des principes ou définitions pratiques des différents « sens » de la responsabilisation;

4. d'élaborer des stratégies permettant d'accroître les connaissances en matière d'algorithme et ainsi de renseigner et former les différents groupes d'intervenants entourant la nature et l'incidence de ceux-ci;
5. d'établir un consensus au sein d'un plus grand bassin de la population en ce qui a trait à ce qui pourrait composer une « trousse de solutions » [21];
6. de mettre au point des indicateurs clés pour la responsabilisation et la justice algorithmique;
7. de clarifier les rôles que jouent les différents acteurs et intervenants dans la responsabilisation en matière d'IA, ou d'en arriver à une entente à ce sujet[22].

Ci-dessous se trouvent des exemples du travail présentement mené dans le but d'élaborer certains principes et normes ainsi que de définir des approches s'appliquant aux différentes compétences.

Déclarations de principes

Les gouvernements et groupes d'intervenants provenant de différents secteurs de niveaux nationaux, régionaux ou municipaux ont fait des déclarations de principes qui guideront différents aspects de l'élaboration, l'acquisition et l'utilisation de l'IA[23]. De plus, un certain nombre d'organisations privées ont mis sur pied des cadres fondés sur des principes en vue d'une adoption responsable de l'IA. Celles-ci comprennent notamment Google, SAP et Microsoft[24].

Japon - La conférence des experts-conseils organisée par le ministère des Affaires intérieures et des Communications du Japon a permis de rédiger une ébauche des principes de recherche et développement en matière d'IA qui fait la promotion des avantages économiques et sociétaux offerts et de la réduction des risques qu'elle représente, tels que le manque de transparence ou la perte de contrôle. La vision d'ensemble de la conférence prône la *société de réseau de sagesse* :

> « […] une société dans laquelle, grâce aux percées faites en matière d'IA, les humains vivent en harmonie avec les réseaux d'IA; où les données, renseignements et connaissances sont créés et distribués librement et de façon sécuritaire en plus d'être reliés à une forme de *réseau de sagesse* favorisant, au-delà de l'espace physique, la collaboration entre les personnes, les choses et les événements provenant de différentes sphères, entraînant conséquemment une créativité et un épanouissement dynamique. » [traduction][25]

Les principes nécessaires à la réalisation de cette vision comprennent la collaboration, la transparence, la contrôlabilité, la sûreté, la sécurité, la protection de la vie privée, l'éthique, l'assistance aux usagers et la responsabilisation.

Faisant fond sur ces travaux, la conférence a permis d'en venir à la rédaction d'une ébauche des principes d'utilisation de l'IA[26], qui les regroupe selon trois piliers, soit la

promotion des avantages, l'atténuation des préjudices et l'établissement de la confiance :

1. Principe d'une utilisation adéquate
2. Principe de la qualité des données
3. Principe de collaboration
4. Principe de sûreté
5. Principe de sécurité
6. Principe de protection de la vie privée
7. Principes de dignité humaine et d'autonomie individuelle
8. Principe d'équité
9. Principe de transparence
10. Principe de responsabilité

Au mois de mai 2018, le Bureau du Cabinet du Japon a entrepris des discussions en vue de formuler les principes sociaux liés à l'IA centrée sur l'humain, qui serviront de fondement à une meilleure implantation sociale et un meilleur partage de l'IA. Les principes sociaux de l'IA seront complétés en mars 2019.

Canada - La déclaration de Montréal pour un développement responsable de l'intelligence artificielle, qui est le résultat d'un processus d'engagement de multiples intervenants et dont l'Université de Montréal est à la tête, vise à souligner « des orientations éthiques pour le développement de l'intelligence artificielle »[27]. La première phase a permis de désigner sept valeurs clés dont il faut tenir compte dans le développement de l'IA : « bien-être, autonomie, justice, vie privée, connaissance, démocratie et responsabilité ».

Élaboration des normes

Plusieurs organisations, soit professionnelles ou autres, travaillent à l'élaboration des normes pour un développement et une utilisation éthiques de l'IA[28].

IEEE (Institute of Electrical and Electronic Engineers) - En 2016, l'IEEE, la plus vaste organisation d'ingénierie professionnelle au monde, a mis sur pied une initiative globale portant sur l'éthique en matière de systèmes autonomes et intelligents (*Global Initiative on Ethics of Autonomous and Intelligent Systems*). Dans sa seconde version, ce document de référence décrit le travail continu qu'accomplissent certains groupes de travail sur les normes, qui ont depuis reçu la tâche d'aborder différents sous-domaines, notamment :

- la confidentialité des données;
- la transparence des systèmes autonomes;

- la création d'un processus modèle permettant de traiter des questions éthiques lors de la conception d'un système;
- des normes favorisant une approche éthique en conception des systèmes robotiques, intelligents et autonomes;
- des paramètres de bien-être pour des systèmes d'IA éthiques et autonomes.

Organisation internationale de normalisation (ISO) - L'Organisation internationale de normalisation (ISO) a récemment créé un sous-comité technique en IA (SC 42), qui travaille à l'élaboration des normes fondamentales ainsi qu'au traitement des enjeux liés à la sécurité et à la fiabilité. SC 42 a mis sur pied des groupes de travail devant se pencher sur les approches informatiques et caractéristiques des systèmes d'intelligence artificielle ainsi que sur leur fiabilité, utilisation et application.

Ces initiatives offrent des points de départ prometteurs et ont le potentiel de mettre à contribution des résultats positifs et significatifs pour chacun des mandats individuels. Il y a beaucoup à apprendre et à retenir de celles-ci. Cependant, il reste beaucoup de travail à faire avant que soit mis au point un ensemble de règles parfaitement formulées, strictes et globales en matière de responsabilisation de l'IA.

Ci-dessous se trouvent des exemples de stratégies formelles entreprises par différentes administrations pouvant servir de précédents à d'autres régions :

Directive sur les processus décisionnels automatisés du gouvernement du Canada

Le gouvernement du Canada travaille présentement à la rédaction de sa toute première Directive sur les processus décisionnels automatisés, qui, dans sa version actuelle[29], formule diverses exigences en matière de conception et d'utilisation de l'IA. Cette directive s'applique uniquement aux systèmes du gouvernement du Canada en développement qui fournissent des services externes, et peut être appliquée à tout système, outil ou modèle utilisé pour prendre des décisions administratives. Ces exigences incluent les règles entourant les évaluations d'impact algorithmique[30]; la transparence et l'explicabilité; le contrôle de la qualité; la garantie d'une intervention humaine; les recours possibles et l'établissement de rapports.

Règlement général sur la protection des données de l'Union européenne (RGPD)

Fidèle à son nom, le RGPD est une initiative de réglementation qui fixe des règles de protection des données générales et qui vise à protéger la vie privée des personnes au sein de l'Union européenne. En plus de préciser les règles entourant le consentement des personnes à l'usage des données personnelles, les actes 13 à 15, tout particulièrement, déterminent ce qui a été décrit comme le « droit à l'explication » lors d'une prise de décision en fonction d'un algorithme, c'est-à-dire le droit des personnes d'exiger certains renseignements qui expliqueraient la logique algorithmique employée dans la prise de décision par un système à l'aide de leurs données personnelles. Certains prétendent que cela poserait des barrières à l'innovation en matière d'IA, liées

à la fois aux coûts directs associés à la révision manuelle des décisions prises par algorithme et aux limites potentielles de performance que cela pourrait entraîner[31], alors que d'autres voient le RGPD comme un pas de l'avant au niveau de la responsabilisation à l'égard de l'IA[32].

Le groupe de travail sur les systèmes décisionnels automatisés de New York

Ce nouveau groupe de travail, considéré comme le premier du genre aux États-Unis, promet de « [recommander] un processus d'examen des systèmes de décision automatisés du gouvernement, plus communément appelés algorithmes »[33]. Il veillera à ce que les algorithmes soient « utilisés de façon appropriée et cadrent avec l'objectif de faire de la ville de New York un endroit plus juste et plus équitable pour tous ses résidents ».

Perspectives d'avenir

S'appuyant sur cet aperçu de l'IA, de la responsabilisation et des activités en cours, la section suivante vise à catalyser la discussion lors de la conférence du 6 décembre et les mesures possibles pour l'avenir. Nous commençons par une courte liste non exhaustive des rôles des groupes d'intervenants potentiels, ainsi que des sujets de discussion et des possibilités de leadership du G7 qui seront envisagés à la conférence.

Rôles des nombreux intervenants

En raison de la complexité et de l'intersectionnalité des questions liées à l'IA et à la responsabilisation, il sera essentiel de créer des occasions inclusives permettant à divers groupes d'intervenants de se réunir pour faire avancer ce travail dans l'intérêt des gens du monde entier. Un certain nombre d'intervenants différents pourraient participer à l'élaboration et au maintien de solides régimes mondiaux de responsabilisation en matière d'IA en fonction des rôles respectifs. Des exemples sont fournis ci-dessous.

Rôles potentiels :

Décideurs au sein des gouvernements nationaux

- Coordonner les activités relatives aux politiques dans des contextes nationaux et internationaux.
- Promouvoir la recherche gouvernementale responsable et inclusive en informatique, en robotique éthique et en génie de l'IA, ainsi qu'en innovation juridique.
- Promouvoir la création et la vérification de la responsabilisation, de la fiabilité et d'autres normes d'IA, tant à l'échelle nationale qu'internationale, qui tiennent compte de l'opacité unique de nombreux systèmes d'IA, et du pouvoir de l'IA d'avoir des répercussions vastes et rapides sur la société.

Organisations intergouvernementales

- Fournir une tribune pour réunir les intervenants internationaux afin de discuter des stratégies de responsabilisation de haut niveau.
- Travailler à l'élaboration de politiques et de mécanismes de coordination pour aborder les questions de responsabilisation et de confiance en matière d'IA.

Décideurs des gouvernements infranationaux

- Convoquer les intervenants locaux pertinents pour élaborer des solutions responsables et inclusives aux défis liés à l'IA dans les administrations infranationales.
- Offrir des occasions d'expérimentation de l'IA responsable et inclusive et partager les pratiques exemplaires avec d'autres administrations.
- Soutenir le travail localisé de responsabilisation et de fiabilité de l'IA qui s'appuie sur le droit, l'économie ou la culture propres à une région.

Sociétés et autres propriétaires de données

- Assurer des niveaux appropriés de contrôle humain dans la conception et l'utilisation de la prise de décisions automatisée (algorithmique).
- Mettre en œuvre des processus d'éthique et de responsabilisation transparents et significatifs tout au long du cycle de vie de l'innovation.
- Définir et promouvoir des codes de conduite à l'appui de la responsabilisation.
- Mobiliser les organismes de réglementation pour les aider à cerner les possibilités de réglementation responsable, par exemple pour aider à coordonner les interventions de l'industrie lorsque les externalités résultent de la prise de décisions algorithmiques.

Universités et collèges

- Veiller à ce que l'éthique de l'IA soit un aspect fondamental des programmes de sciences informatiques et d'ingénierie et à ce que la littératie en codage soit un aspect fondamental des programmes de sciences sociales et des sciences humaines.
- Coordonner la recherche interdisciplinaire, les ateliers et les réunions afin de promouvoir davantage les thèmes de l'ingénierie éthique de la robotique et de l'IA.

Groupes de défense et organismes d'intérêt public

- Fournir des mécanismes de participation citoyenne inclusive.
- Favoriser le dialogue et l'échange de connaissances entre le gouvernement et le secteur privé.
- Promouvoir des ensembles de données équitables et ouverts pour l'apprentissage du modèle.

Fondations

- Investir dans la recherche et l'innovation responsables et inclusives en matière d'IA.
- Fournir des plateformes aux chercheurs pour les aider à créer des trousses d'outils et des cadres d'évaluation de l'IA.

Organismes de réglementation professionnels

- Élaborer des codes de conduite et des mécanismes de responsabilisation pour les membres autorisés, qui tiennent compte de la capacité unique des membres d'avoir des répercussions importantes et rapides sur la société par l'élaboration, l'acquisition, le déploiement et l'utilisation de systèmes d'IA.

Ce que nous avons entendu

Une première version de ce document a été mise en ligne pour consultation publique. Nous avons reçu des commentaires d'un certain nombre de personnes au Canada et au Japon. La plupart de leurs commentaires ont été intégrés au document, mais nous avons également tenté de présenter un résumé ci-dessous. Veuillez noter qu'ils ont été condensés ou reformulés et qu'ils visent à représenter les points de vue des personnes consultées, pas nécessairement ceux des auteurs.

Engagement multipartite

- Veiller à ce que les décideurs politiques discutent avec des experts techniques chevronnés.
- Encourager une véritable diversité dans l'engagement, en particulier auprès des communautés marginalisées et de la société civile.
- Veiller à ce que les conflits d'intérêts soient gérés, en particulier parmi les intervenants qui tireront parti de l'adoption généralisée de l'IA.
- Faire passer la discussion de la responsabilisation à l'éthique en général.
- Tenir compte des diverses intersections de la gouvernance et de la responsabilisation aux niveaux international, national, régional et municipal.
- Veiller à ce que les points de vue des pays à revenu faible ou intermédiaire soient intégrés aux décisions et aux propositions.
- Veiller à ce que la participation ne se limite pas aux bénéficiaires ou aux défenseurs des nouvelles technologies.

Possibilités d'intervention des intervenants

- Promouvoir une plus grande diversité au sein de la main-d'œuvre de la technologie.
- Appuyer la recherche sur l'éducation du public et l'éthique publique.
- Élaborer des normes de données.
- Créer des groupes de travail sectoriels par application.

- Veiller à ce que les processus d'élaboration de règlements et de normes soient ouverts et non restreints par les ressources financières des intervenants.
- Envisager la création d'un comité d'éthique de la recherche ou d'un organisme de style d'essais cliniques pour certaines applications de l'IA.
- Veiller à ce que le personnel du gouvernement à tous les niveaux ait une compréhension suffisante de l'IA pour en assurer la surveillance et cerner les possibilités de modernisation des services.
- Tenir compte des cas où les politiques actuelles, comme la protection de la vie privée, la sécurité, les secrets commerciaux et le droit d'auteur, peuvent créer des obstacles à la responsabilisation.
- Favoriser l'accessibilité au public en diffusant toutes les communications dans un langage simple.
- Intégrer des membres du public à tout organe consultatif ou de gouvernance.
- Mettre au point des mécanismes de soutien au financement afin que les groupes vulnérables et les jeunes puissent participer pleinement aux discussions sur un pied d'égalité avec les représentants de l'industrie.
- Appuyer la promotion de la confiance envers l'IA par l'engagement public des organisations à l'égard de principes précis et, en outre, par la vérification de la conformité par des tiers.
- Bien que le libre marché puisse éventuellement parvenir à un équilibre de l'IA responsable, la période de transition risque de causer des préjudices non négligeables, ce qui justifie la participation du gouvernement.

Autres considérations

- Porter une attention particulière aux scénarios dans lesquels le contrôle humain des systèmes d'IA peut être perdu.
- Examiner les scénarios possibles d'IA, et non seulement les défis et les possibilités actuels
- La transparence est nécessaire, mais pas suffisante.
- Il est essentiel que les nouvelles technologies ne changent pas le rôle fondamental de la société civile.
- Établir des distinctions claires entre les effets de l'IA qui peuvent causer des lésions corporelles et les dommages financiers et traiter chacun de façon appropriée.
- La responsabilisation est avant tout un défi social plutôt que technique.
- Le déploiement de l'IA s'inscrit dans une architecture sociale de contextes culturels, juridiques, économiques et politiques.
- Veiller à ce que les systèmes autonomes rendent des comptes aux personnes touchées par ces systèmes.

- L'accroissement de l'autonomie des systèmes, en particulier l'autonomie physique des robots, augmente divers facteurs de risque.
- Les codes de conduite ne suffiront pas à façonner le comportement et à limiter les abus; des mesures juridiques seront nécessaires.
- Les solutions non techniques, comme les exigences de divulgation pour les grands systèmes commerciaux, peuvent être plus productives que la recherche technique émergente.
- L'industrie de l'IA devra assurer une activité responsable afin de maintenir son acceptabilité sociale, en particulier dans les domaines sensibles, comme les soins de santé.
- Les règles non contraignantes, comme les lignes directrices et les principes de développement et d'utilisation, peuvent devenir de facto des règles impératives, car ces normes sont de plus en plus considérées comme des principes de base pour éviter la négligence.
- La coordination multinationale des normes réduirait les obstacles au déploiement responsable de l'IA.
- Le développement et l'utilisation de l'IA en sont aux premiers stades et, pour ne pas nuire à l'innovation, les principes régissant l'IA devraient être non réglementaires.
- Le dédommagement peut contribuer à promouvoir la responsabilisation.
- Le risque est estimé en multipliant la probabilité par la gravité de la perte présumée. Comme les risques ne sont parfois évalués que par la gravité de la perte, il est nécessaire de bien évaluer le risque en tenant aussi compte de la probabilité qu'il se matérialise.
- Il est possible que les risques évalués varient selon les cultures, et il peut être nécessaire d'établir un niveau de responsabilisation propre à chacune d'elles.
- La préparation d'un mécanisme permettant d'arrêter immédiatement l'utilisation de l'IA lorsque des dommages sont causés par l'utilisation a pour effet d'empêcher la propagation des dommages et d'améliorer la confiance.
- Envisager de promouvoir la fiducie, y compris la création d'un système d'assurance.
- Envisager d'établir une clause d'exemption semblable au comité d'enquête sur les accidents d'aéronef.
- Préciser ce qui doit être expliqué, dans quelle mesure une explication est nécessaire et quel type d'explication est acceptable.

La suite pour le G7

S'appuyant sur les discussions amorcées lors de la réunion ministérielle du G7 sur les TIC à Takamatsu en 2016, les membres du G7 ont entrepris des études sur les enjeux

sociaux, économiques, éthiques et juridiques potentiels soulevés par l'IA, ainsi que sur son incidence socioéconomique.

Le G7 reconnaît également la nécessité de poursuivre le partage de l'information et la discussion afin d'approfondir la compréhension des possibilités et des défis multidimensionnels que présente l'IA. Le G7 et d'autres groupes multilatéraux pourraient jouer un certain nombre de rôles dans la promotion d'une plus grande responsabilisation dans le secteur de l'IA. Des exemples sont fournis ci-dessous.

- Explorer la possibilité d'inclure d'autres membres du G7 dans le nouveau groupe d'étude internationale Canada-France sur l'IA.
- Approuver officiellement une déclaration de principes sur l'IA éthique existante ou en créer une nouvelle.
- Former un groupe de travail du G7 se réunissant régulièrement pour échanger des pratiques exemplaires sur différents sujets, y compris les cadres de responsabilisation et l'utilisation éthique de l'IA au gouvernement.
- S'engager à tenir régulièrement un sommet multilatéral, comme celui du 6 décembre, pour examiner les enjeux émergents de l'IA et de la responsabilisation, dans le cadre d'un forum ouvert et inclusif.
- Engagement du G7 à appuyer les initiatives nationales en matière de responsabilisation.

Conclusion

Nous sommes heureux d'avoir l'occasion de présenter cet aperçu de l'IA et de la responsabilisation dans le but de stimuler un débat rigoureux à la conférence du 6 décembre. À mesure que le développement des applications de l'IA s'élargit et s'accélère, il est urgent et important que les intervenants de tous les secteurs, et au-delà des frontières, se réunissent pour mieux comprendre ce que signifie la responsabilisation dans un monde alimenté par l'IA et les répercussions pour la confiance de la société. Nous espérons que cette enquête sur la responsabilisation en matière d'IA constituera une ressource utile pour les participants à la conférence et les autres, et stimulera la tenue future de discussions parmi les membres du G7, les autres pays et les intervenants du monde entier ainsi que la prise de mesures éventuelles.

---

[1] https://g7.gc.ca/wp-content/uploads/2018/06/FurturInteligenceArtificielle.pdf

[2] WWW Foundation. (2017). « Algorithmic Accountability : Applying the Concept to Different Country Contexts ». *World Wide Web Foundation*. En ligne : https://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf, 5. [traduction]

[3] Sharkey, N. (2018). « The Impact of Gender and Race Bias in AI ». *CICR : Humanitarian Law& Policy.* En ligne : http://blogs.icrc.org/law-and-policy/2018/08/28/impact-gender-race-bias-ai/.

[4] Doshi-Velez, F. et M. Kortz (2017). « Accountability of AI Under the Law : The Role of Explanation ». *Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper.* En ligne : http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584

---

[5] Levin, S. (2018). « Tesla fatal crash : "autopilot" mode sped up car before driver killed, report finds ». *The Guardian*. (8 juin 2018). En ligne : https://www.theguardian.com/technology/2018/jun/07/tesla-fatal-crash-silicon-valley-autopilot-mode-report; Angwin, J. et coll. (2016). « Machine Bias ». *ProPublica.* En ligne : https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[6] Cath et coll. (2017). « Artificial Intelligence and the 'Good Society' : The US, EU, and UK Approach ». *Science and Engineering Ethics 24*(2) : 505-528.

[7] Reisman, D. et coll. (2018). « Algorithmic Impact Assessments : A Practical Framework for Public Agency Accountability ». *AI Now Institute.* En ligne : https://ainowinstitute.org/aiareport2018.pdf; WWW Foundation; Doshi-Velez.; Jones, M. et J. Millar (2017). « Hacking Metaphors in the Anticipatory Governance of Emerging Technology : The Case of Regulating Robots ». Dans R. Brownsword, E. Scotford et K. Yeung (dir.) *The Oxford Handbook on the Law and Regulation of Technology*. (Oxford : Oxford University Press).

[8] Millar, J. (2017). « Ethics Setting for Autonomous Vehicles ». Dans P. Lin, R. Jenkins, K. Abney et G.A. Bekey (dir.) *Robot Ethics 2.0.* (Oxford : Oxford University Press).

[9] Millar, J., et I. Kerr (2016). « Delegation, Relinquishment, Responsibility : The Prospect of Expert Robots ». Dans Ryan Calo, Michael Froomkin et Ian Kerr (dir.) *Robot Law*. (Cheltenham, R.-U. : Edward Elgar Press).

[10] Cowgill, B. (2018). « The Impact of Algorithms on Judicial Discretion : Evidence from Regression Discontinuities ». *Working* Paper. En ligne : http://www.columbia.edu/~bc2656/workingpapers.html

[11] Levin, S. (2018).

[12] Doshi-Velez et Kortz; Villani, C. *Donner un sens à l'intelligence artificielle Pour une stratégie nationale et européenne,* 2018 L'intelligence artificielle au service de l'humain. En ligne : https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf

[13] House of Lords, Select Committee on Artificial Intelligence. 2018. K. Yeung. *AI in the UK: ready, willing and able?,* cité en page 96. En ligne : https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf; Larus, J. et al. 2018. *When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making.* Informatics Europe et EUACM. En ligne : http://www.informatics-europe.org/component/phocadownload/category/10-reports.html?download=74:automated-decision-making-report

[14] Reisman, D. et al. (2018); House of Lords, 35.

[15] Reisman, D. et al. (2018).

[16] World Wide Web (WWW) Foundation

[17] Ibid, 5.

[18] Aitken, M., S. Cunningham-Burley et C. Pagliari. *Moving from Trust to Trustworthiness: Experiences of Public Engagement in the Scottish Health Informatics Programme. Science and Public Policy 43 (5),* 2016, pp. 713–723. Voir aussi Kaminski, M. et al. 2017. *Averting Robot Eyes Maryland Law Review 76* (4):

[19] Future of Humanity Institute. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.* En ligne : https://maliciousaireport.com.

[20] Voir WWW Foundation.

[21] Ibid.

[22] Gilliam, M. (2018). *Governing Artificial Intelligence*, *Data & Society*. En ligne : https://datasociety.net/output/governing-artificial-intelligence/

[23] Veuillez vous référer, par exemple, à : *Draft AI R&D GUIDELINES for International Discussions*, En ligne : http://www.soumu.go.jp/main_content/000507517.pdf

[24] Google AI. *Responsible AI Practices*, https://ai.google/education/responsible-ai-practices; SAP. *SAP's Guiding Principles for Artificial Intelligence*, https://news.sap.com/2018/09/sap-guiding-principles-for-artificial-intelligence/; Microsoft. *Microsoft AI Principles*, https://www.microsoft.com/en-us/ai/our-approach-to-ai.

[25] Ibid.

[26] *Draft AI Utilization Principles*, en ligne : http://www.soumu.go.jp/main_content/000581310.pdf

[27] *Déclaration de Montréal pour un développement responsable de l'intelligence artificielle* en ligne : https://www.declarationmontreal-iaresponsable.com

[28] Veuillez vous référer, par exemple, à : IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*, https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdfet le CIO Strategy Council du Canada : https://ciostrategycouncil.com/standards/new-projects/

---

[29] Une version Google Docs de l'ébauche de la Directive est mise à la disposition du public aux fins de commentaires à l'adresse suivante : https://docs.google.com/document/d/18Tq0piuD03wVr27ZPdQbkq-huZxkIzv7r4IZHt0nezI/edit

[30] Également en cours d'élaboration.

[31] Wallace, N., Castro, D. (2018). « The Impact of the EU's New Data Protection Regulation on AI. » *Center for Data Innovation*. Online: http://www2.datainnovation.org/2018-impact-gdpr-ai.pdf (en anglais seulement)

[32] Doshi-Velez & Kortz.

[33] *New York City Automated Decision Systems Task Force* : https://www1.nyc.gov/site/adstaskforce/index.page (en anglais seulement)