Commission européenne

GROUPE D'EXPERTS DE HAUT NIVEAU SUR L'INTELLIGENCE ARTIFICIELLE

PROJET DE LIGNES DIRECTRICES EN MATIERE D'ETHIQUE POUR UNE IA DIGNE DE CONFIANCE

RESUME

Document de travail pour la consultation des parties prenantes

Bruxelles, le 18 décembre 2018

PROJET DE LIGNES DIRECTRICES EN MATIERE D'ETHIQUE

POUR UNE IA DIGNE DE CONFIANCE

Groupe d'experts de haut niveau sur l'intelligence artificielle
Projet de lignes directrices en matière d'éthique pour une IA digne de confiance

Commission européenne
Direction générale de la communication

Personne de contact Nathalie Smuha – coordinatrice du groupe d'experts de haut niveau sur l'IA
Adresse électronique CNECT-HLG-AI@ec.europa.eu

Commission européenne
B-1049 Bruxelles

Document rendu public le 18 décembre 2018, en anglais.

Le présent document de travail a été élaboré par le groupe d'experts de haut niveau sur l'IA sans préjudice de la position de chacun de ses membres sur des points spécifiques et sans préjudice de la version définitive du document. Le présent document sera encore approfondi et sa version définitive sera présentée en mars 2019 au terme de la consultation des parties prenantes dans le cadre de l'Alliance européenne pour l'IA.

RÉSUMÉ

Le présent document de travail constitue un projet de lignes directrices en matière d'éthique dans le domaine de l'IA, élaborées par le groupe d'experts de haut niveau sur l'intelligence artificielle (GEHN IA) de la Commission européenne. La version définitive de ces lignes directrices est attendue en mars 2019.

L'intelligence artificielle figure parmi les forces les plus transformatrices de notre époque et elle modifiera inéluctablement notre tissu social. Elle constitue une formidable occasion d'accroître la prospérité et la croissance, que l'Europe doit s'efforcer d'atteindre. Ces dix dernières années, des avancées majeures ont été réalisées grâce à la disponibilité de grands volumes de données numériques, à des architectures informatiques puissantes et aux progrès des techniques de l'IA, comme l'apprentissage automatique. L'IA a permis des évolutions importantes dans le domaine des véhicules autonomes, des soins de santé, des robots domestiques et des robots de service, de l'éducation ou de la cybersécurité, qui améliorent la qualité de notre vie au quotidien. En outre, l'IA est essentielle pour relever de nombreux défis majeurs auxquels le monde doit faire face, tels que la santé et le bien-être au niveau mondial, le changement climatique, la fiabilité des systèmes juridiques et démocratiques et d'autres enjeux mentionnés dans les objectifs de développement durable des Nations unies.

Capable de générer des avantages considérables tant pour les particuliers que pour la société, l'IA comporte également certains risques qu'il convient de gérer convenablement. Étant donné que, dans l'ensemble, les avantages de l'IA l'emportent sur les risques qu'elle présente, nous devons veiller à suivre la voie permettant de **maximiser les avantages de l'IA tout en réduisant au minimum les risques.** Afin de rester sur la bonne voie, il est nécessaire de prévoir une **approche de l'IA centrée sur l'humain**, nous obligeant à garder à l'esprit que le développement et l'utilisation de l'IA ne doivent pas être considérés comme une fin en soi, mais bien comme un moyen d'accroître le bien-être humain. **Une IA digne de confiance sera notre étoile polaire,** car les êtres humains ne seront en mesure de tirer pleinement parti des avantages de l'IA en toute sérénité que s'ils peuvent se fier à la technologie.

Une IA digne de confiance présente les **deux caractéristiques suivantes:** 1) elle doit respecter les droits fondamentaux, la réglementation applicable ainsi que les valeurs et principes de base, garantissant une «**finalité éthique**»; et 2) elle doit être fiable et **robuste sur le plan technique** car, même avec de bonnes intentions, un manque de maîtrise technologique peut causer un préjudice involontaire.

Les présentes lignes directrices établissent donc le **cadre d'une IA digne de confiance:**

- le **chapitre I** porte sur la **finalité éthique de l'IA**, en définissant les droits fondamentaux ainsi que les principes et valeurs qu'elle doit respecter;
- sur la base de ces principes, le **chapitre II** donne des **orientations relatives à la mise en œuvre** d'une IA digne de confiance, en abordant les questions d'éthique comme de robustesse technique. Pour ce faire, il dresse une liste des exigences d'une IA digne de confiance et présente une vue d'ensemble des méthodes techniques et non techniques pouvant être utilisées pour sa mise en œuvre;
- le **chapitre III concrétise** alors ces exigences en fournissant une liste d'évaluation concrète mais non exhaustive pour une IA digne de confiance. Cette liste est ensuite adaptée à des cas d'utilisation spécifiques.

À l'inverse d'autres documents traitant de l'IA éthique, les lignes directrices ne visent donc pas à fournir une nouvelle liste de valeurs et de principes fondamentaux pour l'IA, mais plutôt des orientations pour l'application concrète et la mise en œuvre opérationnelle de ces valeurs et de ces principes dans les systèmes d'IA. Ces orientations se présentent sous la forme de trois niveaux d'abstraction, du plus abstrait, au chapitre I (droits fondamentaux, principes et valeurs), au plus concret, au chapitre III (liste d'évaluation).

Les lignes directrices sont destinées à l'ensemble des **parties prenantes concernées qui mettent au point, déploient ou utilisent l'IA**, à savoir des entreprises, des organisations, des chercheurs, des services publics, des institutions, des particuliers ou d'autres entités. La version définitive de ces lignes directrices présentera un mécanisme en vue de permettre aux parties prenantes de les approuver librement.

Il importe de préciser que ces lignes directrices ne visent ni à remplacer toute forme d'élaboration de politiques ou de réglementations (dont il sera question dans la deuxième contribution du GEHN IA: les recommandations en matière de politique et d'investissement, attendues en mai 2019), ni à en décourager l'introduction. Par ailleurs, les lignes directrices devraient être considérées comme un document évolutif devant être mis à jour régulièrement au fil du temps afin d'en maintenir la pertinence, à mesure que la technologie et nos connaissances de celle-ci évoluent. Le présent document devrait donc servir de point de départ à la discussion sur «**Une IA digne de confiance "fabriquée en Europe"**».

Bien que l'Europe ne puisse diffuser son approche éthique de l'IA que si elle est compétitive au niveau mondial, cette **approche éthique de l'IA est essentielle pour permettre une compétitivité responsable**, car elle suscitera la confiance des utilisateurs et facilitera une adoption plus large de l'IA. Les présentes lignes directrices ne visent pas à freiner l'innovation dans le domaine de l'IA en Europe. Leur objectif est, au contraire, d'utiliser l'éthique comme source d'inspiration pour créer une marque unique d'IA, qui protège à la fois les particuliers et le bien commun et leur permet d'en tirer avantage. L'Europe pourra ainsi s'imposer comme leader pour une IA éthique, sûre et de pointe. Ce n'est que si la fiabilité de l'IA est garantie que les citoyens européens pourront bénéficier pleinement de ses avantages.

Enfin, au-delà de l'Europe, les présentes lignes directrices visent à **encourager la réflexion et la discussion** sur un cadre éthique pour l'IA **au niveau mondial**.

ORIENTATIONS

Chaque chapitre des lignes directrices fournit des orientations sur la façon de parvenir à une IA digne de confiance. Celles-ci s'adressent à l'ensemble des parties prenantes concernées qui mettent au point, déploient ou utilisent l'IA, et sont résumées ci-dessous.

Chapitre I: orientation essentielle pour garantir la finalité éthique

- Veiller à ce que l'IA soit **centrée sur l'être humain:** l'IA doit être conçue, déployée et utilisée avec une «**finalité éthique**», fondée sur – et reflétant – les droits fondamentaux, les valeurs sociétales et les principes éthiques de *bienfaisance* (faire le bien), de *non-malfaisance* (ne pas nuire), d'*autonomie des êtres humains*, de *justice* et d'*explicabilité*. Il s'agit d'un aspect essentiel pour parvenir à une **IA digne de confiance.**
- S'appuyer sur les droits fondamentaux, les principes et les valeurs éthiques afin d'évaluer de manière prospective les effets possibles de l'IA sur les êtres humains et le bien commun. Accorder une **attention particulière** aux situations concernant des **groupes plus vulnérables** tels que les enfants, les personnes handicapées ou les minorités, ou aux situations caractérisées par des **asymétries de pouvoir ou d'information**, par exemple entre les employeurs et les travailleurs, ou entre les entreprises et les consommateurs.
- Reconnaître et être conscient que l'IA apporte certes des avantages considérables aux particuliers et à la société, mais qu'elle peut également avoir des incidences négatives. Faire preuve de vigilance en ce qui concerne les domaines les plus préoccupants.

Chapitre II: orientation essentielle pour parvenir à une IA digne de confiance

- Intégrer les **exigences d'une IA digne de confiance dès la première étape de conception:** responsabilisation, gouvernance des données, conception pour tous, gouvernance de l'autonomie de l'IA (supervision humaine), non-discrimination, respect de l'autonomie humaine, respect de la vie privée, robustesse, sécurité, transparence.
- Envisager des méthodes techniques et non techniques afin de garantir la mise en œuvre de ces exigences dans le système d'IA. En outre, garder ces exigences à l'esprit lors de la création de l'équipe chargée de travailler sur le système, du système lui-même, de l'environnement d'essai et des applications potentielles du système.
- Fournir, clairement et de façon proactive, **des informations aux parties prenantes** (clients, travailleurs, etc.) sur les capacités et les limites du système d'IA, afin de leur permettre de formuler des attentes réalistes. Il est essentiel de garantir la **traçabilité** du système d'IA à cet égard.
- **Intégrer** l'IA digne de confiance **dans la culture de l'organisation** et fournir des informations aux parties prenantes sur la manière dont elle est mise en œuvre dans la conception et l'utilisation des systèmes d'IA. L'IA digne de confiance peut également être intégrée dans les chartes de déontologie ou dans les codes de conduite des organisations.
- Garantir la participation et l'**inclusion des parties prenantes** dans la conception et l'élaboration du système d'IA. En outre, garantir une **diversité** lors de la formation des équipes chargées d'élaborer, de mettre en œuvre et de tester le produit.
- S'efforcer de **faciliter la vérifiabilité** des systèmes d'IA, en particulier dans les contextes ou situations critiques. Dans la mesure du possible, concevoir un système permettant de retracer les différentes décisions quant aux divers inputs: données, modèles préformés, etc. Définir par ailleurs des **méthodes d'explication** du système d'IA.
- Garantir un processus spécifique pour la **gouvernance de la responsabilisation.**
- Prévoir **la formation et l'éducation** et veiller à ce que les gestionnaires, les concepteurs, les utilisateurs et les employeurs soient renseignés sur l'IA digne de confiance et formés dans ce domaine.
- Savoir qu'il peut y avoir des tensions fondamentales entre les différents objectifs (la transparence peut déboucher sur de mauvaises utilisations; la détection et la correction de partis pris peuvent être contradiction avec les mesures de protection de la vie privée). Communiquer et documenter ces arbitrages.
- Encourager la recherche et l'innovation en vue de soutenir la mise en œuvre des exigences d'une IA digne de confiance.

Chapitre III: orientation essentielle pour évaluer une IA digne de confiance

- Adopter une **liste d'évaluation** assurant une IA digne de confiance lors de la mise au point, du déploiement ou de l'utilisation d'une IA, et l'adapter au cas d'utilisation spécifique du système.
- Ne pas oublier qu'une liste d'évaluation **ne sera jamais exhaustive** et qu'il ne suffit pas de cocher des cases pour garantir une IA digne de confiance. Il s'agit d'établir un processus continu de détermination des besoins, d'évaluation des solutions et d'amélioration des résultats tout au long du cycle de vie du système d'IA.

Cette orientation s'inscrit dans une vision qui englobe une approche de l'intelligence artificielle centrée sur l'humain, laquelle permettra à l'Europe de devenir un leader mondial de l'innovation pour une IA éthique, sûre et de pointe. Elle s'efforce de faciliter et de permettre le développement d'une «**IA digne de confiance "fabriquée en Europe"**» qui améliorera le bien-être des citoyens européens.