Centre de politique européenne

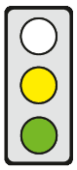
LIGNES DIRECTRICES EN MATIERE D'ETHIQUE POUR L'IA

cep**PolicyBrief** No. 2019/16

ENJEUX-CLES

Objectif de la Communication : La Commission présente des principes éthiques et des exigences essentielles non contraignants pour l'intelligence artificielle (IA), qui doivent être respectés par les développeurs et les utilisateurs de l'IA dans l'ensemble de l'UE.

Parties concernées : développeurs d'IA, entreprises et particuliers qui utilisent ou sont concernés par l'IA.



Pour : (1) Les lignes directrices constituent une première étape utile en direction d'un concept rendant l'IA « acceptable du point de vue éthique » à l'échelle de l'UE.

(2) La « liste d'évaluation pour une IA digne de confiance » contribue à fournir des orientations visant à mettre en œuvre les exigences éthiques.

Contre : Des exemples concrets devraient être ajoutés pour faciliter davantage la mise en œuvre des lignes directrices.

Les passages les plus importants sont indiqués par une ligne dans la marge.

Titre

Communication COM (2019) 168 du 8 avril 2019 : **Renforcer la confiance dans l'intelligence artificielle axée sur le facteur humain,** se référant aux
Lignes directrices en matière d'éthique pour une IA digne de confiance du 8 avril 2019, établies par le « Groupe d'experts de haut niveau sur l'intelligence artificielle »

Note : Les références de page sans autre citation se rapportent à la communication COM(2019) 168, les références « GL» aux « Lignes directrices en matière d'éthique pour une IA de confiance » mentionnées dans la communication (dans leurs versions anglaises).

En bref

► **Contexte général**

– Les systèmes d'intelligence artificielle (« IA ») sont des systèmes logiciels conçus par des êtres humains qui [GL p. 36]
  - appréhendent leur environnement à travers la collecte de données ;
  - interprètent les données collectées ;
  - tirent des conclusions de cette interprétation ou traitent les informations issues des données ;
  - décident des actions les plus appropriées pour atteindre un objectif complexe donné.

– L'IA a le potentiel de transformer notre monde de façon positive : elle peut p. ex. améliorer les soins de santé, réduire la consommation d'énergie, rendre les voitures plus sécurisées et permettre aux agriculteurs d'utiliser les ressources naturelles plus efficacement [p. 1].

– L'intelligence artificielle (« IA ») soulève des questions éthiques et juridiques, notamment parce qu'elle permet à des machines d'« apprendre » de manière indépendante et de prendre des décisions automatisées sans intervention humaine. De telles décisions peuvent être altérées par des cyberattaques ou être erronées, p. ex. elles peuvent être fondées sur des ensembles de données incomplets ou biaisés, et donc non représentatifs [p. 2].

► **Lignes directrices en matière d'éthique pour une « IA digne de confiance »**

– L'IA doit avoir pour objectif d'accroître le bien-être des êtres humains. Par conséquent, l'IA doit être « digne de confiance » et « axée sur le facteur humain », c.-à.-d. « développée de manière à placer les citoyens en son centre » [p. 1, 2].

– À cette fin, la Commission a créé un « groupe d'experts indépendants de haut niveau sur l'IA » (« GEHNIA ») chargé d'élaborer des « Lignes directrices en matière d'éthique pour une IA digne de confiance » [désigné par la suite par les « lignes directrices », p. 2s., GL p. 4].

– Les lignes directrices :
  - sont non contraignantes et visent à créer « des conditions de concurrence équitables sur le plan éthique » au sein de l'UE et à faire de l'UE un leader mondial en matière d'« IA digne de confiance » [p. 2, 9, GL p. 4] ;
  - s'adressent à toutes les parties prenantes, p. ex. les développeurs et les prestataires chargés du déploiement de l'IA – c.-à-d. les utilisateurs privés ou publics de l'IA qui offrent des produits et des services à autrui –, mais aussi aux utilisateurs finaux et à d'autres personnes directement ou indirectement concernées par l'IA [GL p. 5, 14].

– Les lignes directrices prévoient que, pour être « digne de confiance », l'IA soit tout au long de son cycle de vie [p. 3, GL p. 5-10] :
  - licite, c.-à-d. qu'elle respecte la législation applicable et en particulier les droits fondamentaux juridiquement contraignants ;
  - robuste, en particulier sur le plan technique, p. ex. qu'elle soit résiliente aux cyberattaques ;
  - éthique, c.-à-d. qu'elle garantisse l'adhésion aux valeurs éthiques, même lorsque la législation ne suit pas le rythme des progrès techniques. Bien que non contraignantes, ces valeurs éthiques contribuent à identifier :
  - ce qui devrait plutôt que ce qui pourrait être fait avec l'IA ;
  - comment l'IA peut mettre en jeu les droits fondamentaux et leurs valeurs sous-jacentes [GL p. 10].

Auteurs : Alessandro Gasparotti et Dr. Anja Hoffmann, LL.M. Eur. | hoffmann@cep.eu
cepFrance | 350, rue Lecourbe | 75015 Paris | France | Téléphone : +33 (0)1 45 54 91 50 | www.cepfrance.eu

– Les lignes directrices ne traitent pas de l'IA « licite », mais visent à fournir des orientations pour la création d'une IA « éthique » et « robuste ». Cependant, les droits fondamentaux sont également pertinents pour « l'IA éthique » : les droits fondamentaux comprennent des valeurs sous-jacentes et sont conférés aux individus en vertu de leur statut moral en tant qu'êtres humains. [GL p. 6, 10]
– Les lignes directrices [GL p. 9, 14s.] :
  - présentent les « fondements » d'une « IA digne de confiance », qui reposent sur les droits fondamentaux et engendrent quatre « principes éthiques » ;
  - traduisent ces principes théoriques en sept exigences essentielles pour une « IA digne de confiance ».

► **Les droits fondamentaux et les principes éthiques, « fondements » d'une « IA digne de confiance »**

– Une « IA digne de confiance » doit être fondée sur les droits et valeurs fondamentaux consacrés dans les traités de l'UE, la Charte des droits fondamentaux de l'UE et le droit international [p. 2, GL p. 9, 10]. Les droits fondamentaux comprennent notamment les droits suivants, particulièrement importants pour l'IA :
  - le respect de la dignité humaine, ce qui suppose que chaque être humain possède une « valeur intrinsèque » ;
  - la liberté des individus, p. ex. la liberté d'expression et de réunion, le droit à la vie privée et à la confidentialité ;
  - l'égalité, la non-discrimination et la solidarité.
– L'IA doit respecter quatre principes éthiques. Ces principes reflètent les droits fondamentaux [GL p. 11-13] :
  - le respect de l'autonomie humaine : les êtres humains doivent conserver leur autodétermination totale et un contrôle sur l'IA ; l'IA ne doit pas tromper ou manipuler les êtres humains sans justification ;
  - la prévention de toute atteinte : l'IA ne devrait ni porter atteinte ni nuire d'une quelconque autre manière aux êtres humains;
  - l'équité : les bénéfices et les coûts de l'IA doivent être répartis équitablement au sein de la société, la discrimination et les biais injustes doivent être évités et des mécanismes de recours efficaces contre les décisions prises par l'IA devraient être disponibles ;
  - l'explicabilité : l'objectif de l'IA doit être communiqué, les processus transparents et les décisions doivent pouvoir être expliquées autant que possible, en fonction du contexte et de la gravité des conséquences d'une décision erronée.
– Ces principes peuvent être en conflit, p. ex. la technologie de reconnaissance faciale peut réduire la criminalité (prévention de toute atteinte) tout en limitant la vie privée et la liberté individuelle (c.-à-d. l'autonomie humaine).
– Ces arbitrages devraient être reconnus et rendus à l'aide d'une « réflexion raisonnée ». La dignité humaine ne peut toutefois pas être mise en balance avec d'autres droits [GL p. 13].
– Les lignes directrices traduisent ces principes éthiques en une liste non exhaustive de sept exigences essentielles que l'« IA digne de confiance » devrait respecter [p. 4-6, GL p. 15-20] :

► **Sept exigences essentielles pour la réalisation d'une « IA digne de confiance »**

– **Action et contrôle humains** : Afin de préserver l'autonomie et la prise de décision humaines, l'IA devrait soutenir l'action humaine et permettre un contrôle humain [p. 4, GL p. 15, 16] :
  - L'IA devrait aider les êtres humains à prendre de meilleures décisions et faire des choix plus éclairés, et soutenir ainsi leur action.
  - L'IA devrait permettre un contrôle humain qui peut être atteint par le biais de mécanismes de gouvernance ; p. ex.
    - garantir des marges d'appréciation pour les interventions humaines et la capacité d'ignorer des décisions fondées sur l'IA et de décider de ne pas utiliser l'IA dans des situations données ;
    - surveiller l'activité de l'IA ; moins il est possible de contrôler l'IA, plus il faut approfondir les tests.
  - Les autorités publiques doivent pouvoir exercer un contrôle sur l'utilisation des systèmes d'IA conformément à leur mandat.
  - Les effets négatifs sur les droits fondamentaux devraient être évalués avant le développement de l'IA et doivent être réduits ou justifiés.
– **Robustesse technique et sécurité** : Pour prévenir toute atteinte, l'IA doit être techniquement robuste et sécurisée, c.-à-d. [p. 5, GL p. 16, 17] :
  - avoir un niveau de sécurité proportionné à l'ampleur du risque posé par un système d'IA ;
  - être fiable, sécurisée et résistante aux attaques, p. ex. le piratage et la manipulation ;
  - créer les mêmes résultats, c.-à-d. des résultats reproductibles dans les mêmes conditions, de sorte que son comportement puisse être décrit ;
  - être précis et informer les utilisateurs sur la probabilité d'erreurs (p. ex. une précision limitée) ;
  - présenter des garanties permettant le déclenchement d'un plan de secours en cas de problème, p. ex. impliquer un opérateur humain.
– **Respect de la vie privée et gouvernance des données :** une gouvernance appropriée des données est nécessaire pour prévenir les atteintes à la vie privée [p. 5, GL p. 17] :
  - la qualité, l'intégrité et la pertinence des données introduites dans un système d'IA doivent être garanties pour éviter des biais et des erreurs ;
  - les êtres humains doivent avoir un contrôle total sur les données les concernant qui sont collectées pour ou par l'IA ;
  - ces données ne doivent pas être utilisées illégalement.
– **Transparence :** Pour faire en sorte que l'IA soit « explicable », ses éléments – données, systèmes et modèles économiques
  - doivent être transparents ; en particulier [p. 5, GL p. 18] :
  - les systèmes d'IA doivent être traçables, p. ex. en documentant leurs décisions et le processus sous-jacent (y compris les données).

- Les processus décisionnels algorithmiques doivent pouvoir être expliqués autant que possible et des arbitrages doivent être effectués entre leur explicabilité et une réduction potentielle de leur précision. En outre, l'explicabilité est conditionnée au fait que l'IA en question ait « une incidence importante sur la vie des personnes ».
- Les utilisateurs doivent être conscients qu'ils interagissent avec un système d'IA, être informés des limites du système et pouvoir demander une interaction humaine « afin de garantir le respect des droits fondamentaux ».

– **Diversité, non-discrimination et équité** : pour être équitable, l'IA doit être conçue de manière à permettre à chacun d'accéder de manière égale au produit ou au service. Les parties prenantes concernées doivent être impliquées dans les processus de conception. Les biais « injustes » doivent être évités, p. ex. dans les ensembles de données, car ils pourraient conduire à des discriminations. [p. 6, GL p. 18]
– **Bien-être sociétal et environnemental** : pour être équitable et prévenir toute atteinte, l'IA devrait être durable, respectueuse de l'environnement et de la société ; ses incidences sur l'environnement, les êtres humains, la société et la démocratie doivent être surveillés [p. 6].
– **Responsabilité** : pour être équitable, l'IA devrait être conçue de manière à pouvoir être vérifiée par des auditeurs – sans qu'il soit nécessaire de divulguer des informations relatives à la propriété intellectuelle ou autres informations confidentielles –, notamment lorsque son utilisation a des incidences sur les droits fondamentaux. Les effets négatifs devraient être signalés, minimisés et les mécanismes de recours adaptés devraient être prévus. [p. 6, GL p. 20]
– Les exigences peuvent être mises en œuvre à l'aide [GL p. 20-23] :
  - de « méthodes techniques », p. ex. la mise en œuvre de procédures que l'IA doit suivre ou ne pas suivre, et l'« éthique dès la conception », c.-à-d. veiller au respect des normes éthiques dès le début du processus de conception de l'IA et/ou
  - de « méthodes non techniques » de gouvernance, telles que la réglementation, les codes de conduite, la normalisation et la certification.

► **Liste d'évaluation, phase pilote et recherche de consensus international**

– Pour garantir la mise en œuvre pratique des lignes directrices, le GEHNIA a concrétisé les exigences dans une « liste d'évaluation pour une IA digne de confiance », censée aider principalement les développeurs et les prestataires chargés du déploiement de l'IA à atteindre une « IA digne de confiance » [GL p. 25-31].
– Pendant la phase pilote jusqu'en décembre 2019, les parties prenantes peuvent tester la liste et formuler des commentaires. Le GEHNIA mettra à jour les lignes directrices début 2020 [p. 7, GL p. 24].

Contexte politique

En 2018, la Commission a publié une stratégie en matière d'IA [COM(2018) 237] et un « Plan coordonné » [COM(2018) 795], qui ont notamment pour objectif de garantir des règles juridiques et éthiques appropriées pour l'IA [cf. cep**PolicyBrief** n° 2019/13, voir aussi cep**PolicyBriefs** n° 2019/10 et n° 2019/12]. En 2017, le Parlement européen a proposé un code de conduite éthique [résolution PE] et a demandé, en 2019, un cadre éthique pour « une IA centrée sur l'homme » [résolution PE]. Le Conseil a souligné l'importance de la mise en œuvre de lignes directrices en matière d'éthique pour l'IA dans l'Union européenne et au niveau mondial [Conclusions de 02/2019]. En mai 2019, l'OCDE a publié des lignes directrices en matière d'éthique dans le domaine de l'IA, approuvées par le G20 [cf. cep**Input** n° 2019/07].

Options pour influencer le processus politique

Direction générale :	DG Réseaux de communication, contenu et technologies
Commission du Parlement européen :	Industrie, recherche et énergie (principale)

EVALUATION

Évaluation de l'impact économique

L'objectif de la Commission de promouvoir une « IA digne de confiance » peut faciliter l'acceptation de cette technologie. Cependant, **les lignes directrices sont trop générales et imprécises pour être mises en œuvre directement**. En outre, elles ne permettent pas d'accroître la sécurité juridique car elles n'aident pas les développeurs d'IA à respecter la législation applicable. On peut donc se demander si seules les lignes directrices établiront des conditions de concurrence équitables sur le plan éthique dans le domaine de l'IA à l'échelle de l'UE et feront ainsi de l'UE un chef de file mondial en ce qui concerne l'IA digne de confiance. **Toutefois, les** sept **exigences essentielles fournissent un cadre général pour le développement futur de lignes directrices plus précises** – p. ex. sectorielles.

En outre, lorsque des tensions existent entre les principes éthiques lors du développement et de l'utilisation de l'IA, le public devrait également être informé de la manière dont le développeur a effectué l'arbitrage en matière d'éthique. Ces informations ne nuisent pas aux entreprises et ne leur imposent pas une charge importante à quelque niveau que ce soit. Les lignes directrices devraient offrir des solutions ou des exemples sur la manière de rendre de tels arbitrages en matière d'éthique.

Les sept exigences peuvent être évaluées ainsi :

(1) Il convient effectivement de soumettre l'IA à un contrôle humain et de garantir des marges d'appréciation pour les interventions humaines, p. ex. pour ignorer des décisions prises par l'IA, biaisées ou manipulées. Cependant, aucune définition du degré d'appréciation approprié n'est fournie.

(2) Il convient effectivement pour l'IA de déployer un niveau de sécurité proportionné à l'ampleur de son risque. Il serait ainsi disproportionné d'utiliser les mêmes normes pour un réacteur nucléaire contrôlé par l'IA que pour fournir aux utilisateurs des suggestions de titres musicaux. Cependant, en ne définissant aucune norme, les lignes directrices incitent les entreprises à minimiser les risques potentiels liés à leur IA pour respecter leurs propres normes minimales.

(3) La qualité, l'intégrité et la pertinence des données introduites dans un système d'IA sont essentielles pour réduire le risque de résultats non fiables. Les marchés concurrentiels tendent à garantir ce résultat. Toutefois, des exemples devraient être inclus pour fournir des orientations sur le traitement des données.
(4) Afin de favoriser la transparence, les décisions prises par l'IA doivent être traçables et explicables autant que possible. Les utilisateurs bien informés sont mieux placés pour évaluer dans quelle mesure ils acceptent de s'en remettre à l'IA. Il devrait en particulier être possible pour les utilisateurs de voir et de corriger les informations les concernant, telles que leur âge ou leurs intérêts, et qui sont collectées par l'IA et utilisées dans la prise de décision.
(5) Pour garantir la non-discrimination, il conviendrait de clarifier les cas où un biais est « injuste ». Il est peu probable qu'il soit possible d'impliquer toutes les parties prenantes concernées dans les processus de conception. Cependant, les fabricants ont tout intérêt à inclure pleinement celles-ci, car les retours d'expériences améliorent la qualité de l'IA.
(6) La signification de « bien-être sociétal » n'est pas claire. Par exemple, la « liste d'évaluation pour une IA digne de confiance » indique que, pour être respectueuse de la société, l'IA doit contrer le risque de pertes d'emplois. Il est difficile de savoir si en remplaçant les responsables des ressources humaines et en effectuant des entretiens plus équitables, par exemple non discriminatoires, l'IA peut être considérée comme un moyen d'accroître le bien-être sociétal, étant donné que certains responsables des RH perdraient leur emploi.
(7) Afin de garantir la responsabilité de l'IA sans réduire les incitations à innover, l'exigence d'auditabilité devrait autant que possible éviter la divulgation d'informations confidentielles. Outre cet aspect, des mécanismes de recours adaptés sont essentiels pour renforcer la sécurité juridique aussi bien pour les entreprises que pour les consommateurs.

La « liste d'évaluation pour une IA digne de confiance » permet de fournir des orientations sur la mise en œuvre des exigences éthiques. Cependant, **des exemples concrets devraient être ajoutés pour faciliter davantage la mise en œuvre des lignes directrices.** Outre le test pratique – utile – de la liste, effectué par les développeurs et les prestataires d'IA, il conviendrait de favoriser activement un débat public général et transparent sur l'IA, dans lequel les organismes du domaine de l'éthique, les universités, les parlements et le public doivent être intégrés plus activement. Dans le même temps, il conviendrait d'aborder en détail la protection – éventuellement au-delà des lignes directrices – contre les risques futurs et la mise en place éventuelle de limites éthiques à l'utilisation de l'IA dans l'UE.

Évaluation juridique

Compétence

Cela ne pose pas de problème. Il ne s'agit pas de lignes directrices juridiques mais « éthiques » et, par ailleurs, elles ne sont pas contraignantes.

Subsidiarité

Cela ne pose pas de problème (voir ci-dessus).

Proportionnalité à l'égard des États membres

Cela dépend de la conception des mesures de suivi.

Compatibilité avec le droit de l'UE à d'autres égards

Une IA « digne de confiance » doit respecter la législation et en particulier les droits fondamentaux. Cependant, étant donné que les lignes directrices contiennent des exigences applicables à l'IA « éthiques » plutôt que juridiques, il n'y a aucun doute quant à leur compatibilité avec le droit de l'UE. Au lieu de cela, la question – qui nécessite d'être examinée de toute urgence – est de savoir si et dans quelles circonstances l'utilisation croissante de l'IA est acceptable du point de vue éthique ou même nécessaire.

Les lignes directrices constituent une première étape utile en direction d'un concept rendant l'IA « acceptable du point de vue éthique » à l'échelle de l'UE. Les points de vue éthiques, les valeurs et les attitudes morales varient dans l'UE en dépit de valeurs fondamentales communes, créant ainsi un risque que les exigences éthiques applicables à l'IA soient fragmentées. Les quatre principes éthiques retenus – autonomie humaine, prévention de toute atteinte, équité et explicabilité de l'IA – constituent un point de départ adéquat. Ils découlent à juste titre des droits et valeurs fondamentaux qui s'appliquent dans l'UE, et en particulier de la garantie de la dignité humaine, qui est la valeur fondamentale première [art. 2, phrase 1 TUE] et résulte des traditions constitutionnelles communes aux États membres [Meyer-Borowski, Charte des droits fondamentaux de l'UE, art. 1 §26]. Ainsi, dans les questions d'interprétation, il existe un lien avec la magistrature de niveau constitutionnel, et il est possible d'éviter d'importantes divergences entre l'éthique et la loi. Etant donné que le socle de valeurs de l'UE place les êtres humains au centre du projet européen [Calliess / Ruffert, art. 2 VUE, §11], la décision de l'UE d'opter pour une IA « axée sur le facteur humain » est logique.

Conclusion

Les lignes directrices sont trop générales et imprécises pour être mises en œuvre directement. Toutefois, les exigences essentielles fournissent un cadre général pour le développement futur de lignes directrices plus précises. Les lignes directrices constituent donc une première étape utile en direction d'un concept rendant l'IA « acceptable du point de vue éthique » à l'échelle de l'UE. La « liste d'évaluation pour une IA digne de confiance » permet de fournir des orientations sur la mise en œuvre des exigences éthiques. Cependant, des exemples concrets devraient être ajoutés pour faciliter davantage la mise en œuvre des lignes directrices.