Charte d'engagement pour une intelligence artificielle plus sûre pour les enfants

Coalition pour les droits et la protection des enfants à l'ère de l'intelligence artificielle

Préambule

L'intelligence artificielle générative fait son entrée dans des environnements essentiels au développement des enfants, notamment l'éducation, le jeu, la communication, la recherche d'informations, la création, le soutien émotionnel et la vie familiale. À terme, cette technologie pourrait influencer la manière dont les jeunes se développent, apprennent, nouent des relations, participent à la société et y contribuent.

L'IA pourrait offrir des possibilités considérables en matière d'apprentissage, de créativité, d'accessibilité, de communication et d'accompagnement, ainsi que pour prévenir, détecter et combattre les violences et l'exploitation sexuelles commises sur des enfants. Ces avantages doivent être mis en regard du fait que ces systèmes peuvent s'exprimer avec aisance, être profondément personnalisés, adaptatifs et multimodaux, et interagir en répondant aux signaux sociaux ou émotionnels, ce qui comporte un risque d'effets négatifs sur les enfants. L'objectif est de tirer parti des bénéfices de l'IA générative tout en constituant les données probantes, la compréhension commune et les dispositifs de protection nécessaires pour atténuer les risques pesant sur le développement social et cognitif, avant que les habitudes d'utilisation ne deviennent plus difficiles à modifier.

Les risques et les possibilités de l'IA générative doivent être évalués selon une approche différenciée en fonction du développement de l'enfant, en examinant les usages, les fonctionnalités et les modalités d'interaction selon les tranches d'âge, les contextes de développement et les niveaux d'utilisation.

Les connaissances scientifiques sur les interactions répétées des enfants et des adolescents avec l'IA générative sont encore en cours de constitution, et l'on ne dispose pas encore d'études longitudinales suffisamment abouties. Pourtant, l'adoption de ces technologies à grande échelle est déjà engagée, et des décisions relatives aux produits et aux politiques publiques sont déjà prises.

Dans ce contexte, les sciences du développement existantes, les travaux connexes, les premiers résultats empiriques, la connaissance des produits et l'expertise constituent les meilleures bases disponibles pour agir. Ces connaissances doivent toutefois être structurées autour des priorités auxquelles les parties prenantes sont déjà confrontées. Si des mesures de protection doivent être mises en œuvre dès maintenant, proportionnellement aux risques identifiés, elles devront être régulièrement affinées et actualisées à mesure que les lacunes dans les connaissances seront comblées. La Coalition pour les droits et la protection des enfants à l'ère de l'IA, conduite par la France, l'Espagne et le Kenya, et associant dix organismes des Nations unies, réunit les gouvernements, les chercheurs, la société civile et les entreprises pour combler ces lacunes. La présente Charte vise à renforcer les connaissances scientifiques indépendantes sur les risques et les possibilités de l'IA pour le développement des enfants, et à intégrer ces connaissances à la conception des produits, à leur évaluation et aux politiques publiques.

La présente Charte constitue la contribution de ses partenaires industriels à un dialogue multipartite plus large associant les gouvernements, le monde universitaire, les chercheurs, la société civile et d'autres parties prenantes engagées en faveur d'une IA sûre et adaptée à l'âge des enfants. Elle s'appuie sur un ensemble croissant d'engagements politiques et normatifs pris par ses membres[1].

Principes communs

La Coalition est guidée par les principes suivants, destinés à orienter la conception, la mise en œuvre, le fonctionnement et l'atténuation des risques des systèmes d'IA auxquels les enfants peuvent accéder :

Sécurité dès la conception et par défaut : Prendre des mesures concrètes pour garantir aux enfants un niveau élevé de sûreté, de protection de la vie privée et de sécurité dès la conception et par défaut, en appliquant les mesures d'atténuation pertinentes pour protéger leur sécurité et leur bien-être. Ces mesures peuvent notamment comprendre des outils de gestion du temps adaptés à l'âge et renforçant l'autonomie, afin de prévenir les usages excessifs ou compulsifs et les services favorisant l'attachement émotionnel.

Une IA donnant la priorité à l'intérêt supérieur de l'enfant : Respecter les droits des enfants et favoriser des solutions d'IA élaborées selon des approches centrées sur l'enfant, donnant la priorité à ses besoins de développement et préservant des temps hors ligne. Cela comprend des pratiques transparentes et un contrôle parental approprié afin d'aligner les systèmes d'IA sur les droits, les besoins et le bien-être des enfants. Cela implique également de soutenir la recherche internationale sur l'anthropomorphisme, la dépendance excessive et la dépendance émotionnelle.

Une IA éthique, équitable et inclusive : Mettre en œuvre une IA visant à bénéficier à tous les enfants, quels que soient leur origine, leurs capacités ou leurs besoins, en réduisant la fracture numérique. Les systèmes d'IA doivent constamment chercher à favoriser l'inclusion et à garantir l'égalité d'accès.

Une IA adaptée à l'âge dès la conception : Concevoir et utiliser des solutions d'IA adaptées aux stades de développement cognitif et socio-émotionnel des enfants, en tenant compte des différences de maturité, de vulnérabilité et de développement, ainsi que de leurs droits à l'information et à la participation.

Une conception associant les familles : Fournir aux parents et aux responsables légaux des outils efficaces, accessibles et faciles à utiliser, interopérables lorsque cela est techniquement possible, ainsi que des ressources d'accompagnement leur permettant de gérer en confiance les expériences de leurs enfants avec l'IA. Cela suppose également de prendre en compte les points de vue, les besoins et les expériences des enfants pour éclairer l'approche des entreprises dans ce domaine.

Une participation effective des enfants : Soutenir une participation réelle, adaptée à l'âge et éthique des enfants et des jeunes aux travaux de recherche, afin que leurs points de vue, leurs expériences et leurs besoins contribuent à l'élaboration de recommandations, de bonnes pratiques et d'approches relatives aux systèmes d'IA qui les concernent.

Protection de la vie privée des enfants : Protéger les données personnelles des enfants au moyen de garanties solides et de la minimisation des données ; expliquer, de manière adaptée à leur âge, comment les données sont collectées, utilisées et protégées ; et assurer une transparence tenant compte des intérêts propres aux enfants.

Vérification de l'âge efficace et respectueuse de la vie privée : Mettre en œuvre des mécanismes de vérification de l'âge robustes, fiables, précis, fondés sur les risques, respectueux de la vie privée, non discriminatoires et interopérables, ne reposant pas uniquement sur l'autodéclaration, et soutenir l'évaluation indépendante de la précision, du respect de la vie privée et de l'équité de ces solutions.

Combattre et prévenir l'utilisation, dans le développement des systèmes d'IA, de contenus pédocriminels ou d'images intimes diffusées sans consentement, et interdire ces contenus, qu'ils soient ou non générés par l'IA : Interdire l'utilisation des systèmes d'IA à des fins de production ou d'exploitation de contenus montrant des violences sexuelles sur enfants, d'images intimes non consenties ou de tout comportement constituant une exploitation ou une violence sexuelle à l'encontre d'enfants. Concevoir délibérément des systèmes d'IA qui ne facilitent pas les violences et l'exploitation sexuelles des enfants, tout en restant capables de répondre efficacement aux révélations de telles violences ; mettre en œuvre des mesures de prévention et de détection et soutenir leur évaluation, conformément au droit applicable et dans le respect des droits des victimes.

Culture de l'IA et capacité d'agir : Promouvoir une compréhension de l'IA adaptée à l'âge afin que les enfants et les jeunes puissent comprendre le fonctionnement des systèmes d'IA, faire des choix éclairés, reconnaître les risques et les possibilités, et participer effectivement aux décisions qui concernent leur environnement numérique.

Un engagement commun en faveur de la recherche indépendante sur la sécurité des jeunes face à l'IA et les possibilités qu'elle offre

Pour mettre ces principes en pratique, la Coalition pour les droits et la protection des enfants à l'ère de l'IA établira un réseau de recherche scientifique et travaillera avec les gouvernements participants, les institutions publiques, les entreprises privées et d'autres partenaires tiers. Placé sous l'orientation d'un Comité scientifique indépendant, ce réseau reliera les centres de recherche publics, universitaires et privés existants, les instituts de sécurité de l'IA et, le cas échéant, leur réseau international, ainsi que les organisations de la société civile, en mobilisant également les réseaux et l'expertise des États membres et des entités des Nations unies participant à la Coalition. Il aura pour objet de renforcer les connaissances indépendantes sur les risques et les possibilités de l'IA pour le développement des enfants, d'élaborer des outils de mesure communs et de mettre ces connaissances à la disposition du public pour la conception des produits, leur évaluation et les politiques publiques.

Le Comité scientifique s'efforcera de réunir des experts issus de régions géographiques diverses et de garantir une représentation équilibrée des femmes et des hommes, afin que ses recherches, recommandations et bonnes pratiques reflètent les besoins et les réalités des enfants du monde entier.

La Coalition pour les droits et la protection des enfants à l'ère de l'IA facilitera la participation et la réalisation des travaux dans l'ensemble du programme de recherche (annexe 1).

Participation et mise en œuvre

Le programme de recherche sera mis en œuvre grâce aux contributions convenues avec les partenaires et coordonné par l'organisation indépendante mandatée par la Coalition. Les dispositions suivantes régissent ces contributions, la publication des résultats, l'évaluation indépendante et les garanties applicables à la recherche.

[Contributions] Les entreprises participantes et les autres partenaires soutiennent cet effort commun par des contributions financières, une expertise scientifique et technique, un accès aux modèles et aux données pertinentes lorsque la loi le permet, des ressources en nature et des partenariats locaux, en fonction de leurs rôles, de leurs capacités, de leur taille et de l'utilisation de leurs services par les enfants. L'objet, les ressources et le calendrier de réalisation de chaque contribution sont rendus publics. Les contributions sont apportées dans le cadre de la Coalition et coordonnées par l'organisation indépendante afin de soutenir les résultats attendus du réseau de recherche (annexe 1) et ses travaux plus larges. Les contributeurs rendront compte de la mise en œuvre et de toute modification substantielle. Les contributions financières respecteront les règles de financement du programme et les garanties d'indépendance de la recherche.

[Publication] Les méthodes de recherche, les outils de mesure communs, les méthodes d'évaluation indépendante, les approches et résultats des référentiels d'évaluation, y compris les résultats nuls, négatifs ou non concluants, seront publiés. Les équipes de recherche conservent la maîtrise de leurs conclusions, sous réserve de garanties appropriées concernant les informations personnelles et confidentielles. Les résultats seront diffusés sous des formes utiles aux gouvernements, aux auditeurs, aux organismes de normalisation, aux entreprises, aux chercheurs, aux éducateurs, aux familles et aux jeunes.

[Qualité et communication des résultats de la recherche et des évaluations] Les recherches et évaluations indépendantes élaborées ou soutenues dans le cadre de la présente Charte devraient produire des données fiables et pertinentes au regard du développement de l'enfant sur les risques encourus par les enfants et l'efficacité des mesures de protection.

[Audit et transparence méthodologique] Les entreprises participantes ont le droit d'auditer les méthodes d'évaluation appliquées à leurs systèmes et de soulever des préoccupations méthodologiques documentées. Les méthodes d'évaluation doivent satisfaire aux mêmes exigences de validité, de transparence, de publication des résultats et d'examen indépendant que les travaux de recherche. Les protocoles, critères et seuils, la sélection des tests, les méthodes de notation et d'analyse, les limites et les résultats doivent être publiés avec suffisamment de

détails pour permettre un examen indépendant et, lorsque cela est possible, une reproduction des travaux, sous réserve de garanties appropriées de confidentialité et de sécurité. Les conclusions des audits et les réponses des évaluateurs doivent être documentées. Les droits d'audit ne confèrent aucun pouvoir sur les critères de sécurité, les conclusions ou la publication, ni aucun droit de retarder ou d'empêcher la diffusion des résultats.

[Qualité et communication des résultats des évaluations] Les évaluations élaborées ou soutenues dans le cadre de la présente Charte devraient produire des données fiables et pertinentes au regard du développement de l'enfant sur les risques encourus par les enfants et l'efficacité des mesures de protection. Les méthodes d'évaluation et la communication des résultats devraient suivre la méthode et les exigences définies à l'annexe 2.

[Évaluation indépendante] Outre l'évaluation et le suivi de leurs propres systèmes, les entreprises participantes et les autres partenaires soutiendront une collaboration avec l'écosystème des évaluateurs indépendants, notamment les instituts d'évaluation soutenus par les pouvoirs publics, les instituts de recherche et les spécialistes de la protection de l'enfance, afin d'élaborer des méthodologies de recherche et des approches d'évaluation communes concernant les risques spécifiques que les systèmes d'IA présentent pour la sécurité des enfants. Dans le cadre de l'évaluation indépendante, les priorités de recherche, les méthodes, les analyses et les conclusions demeurent sous responsabilité scientifique indépendante. Les entreprises participantes et les autres partenaires travailleront également à des méthodologies communes d'évaluation et pourront fournir aux évaluateurs un accès approprié aux systèmes concernés, afin que les effets des systèmes d'IA sur la sécurité et le développement des enfants, ainsi que l'efficacité des interventions de protection, puissent être évalués objectivement au moyen de méthodes transparentes éclairées par le programme de recherche.

Tout au long du programme de recherche, les priorités, les méthodes, les analyses et les conclusions demeurent sous responsabilité scientifique indépendante. Les évaluateurs indépendants restent responsables des méthodes d'évaluation, de la notation et des conclusions, en appliquant les critères de sécurité et d'adéquation définis de manière indépendante et décrits ci-dessus.

Le soutien financier ou en nature ne confère aucun contrôle sur les décisions individuelles de financement ou sur les résultats, ni aucun droit de veto sur leur publication. Les intérêts pertinents doivent être déclarés et gérés, et la participation des entreprises ainsi que les sources de financement doivent être rendues transparentes.

[Organisation du financement] Les financements industriels soutiennent le volet « Mesures communes ». L'Observatoire, le volet « Données probantes et effets sur le développement », la rémunération du Comité scientifique ainsi que le volet « Évaluation et normes » seront financés exclusivement, directement ou indirectement, par des fonds publics et des sources philanthropiques indépendantes. Une comptabilité distincte garantira le respect de cette séparation.

[Garanties] Les droits des enfants, leur protection, le consentement et la protection des données s'appliquent à l'ensemble du programme. L'accès aux modèles et aux données sera établi selon des modalités convenues entre toutes les parties, protégeant les participants et permettant l'analyse et la publication indépendantes. Les contributions ne donnent aucun accès privilégié aux données personnelles des enfants. Les conditions de participation préserveront ces garanties tout en permettant une coopération durable entre la recherche, les entreprises et les partenaires œuvrant dans l'intérêt général.

Un engagement commun en faveur de l'évaluation des risques, des tests et de la réponse aux incidents

Les entreprises participantes qui développent ou déploient des systèmes d'IA accessibles aux mineurs procéderont, avant leur déploiement lorsque cela est pertinent, puis à intervalles réguliers, à des évaluations et à des tests rigoureux de ces systèmes afin d'identifier les risques pour la santé, la sécurité, les droits fondamentaux ainsi que le bien-être et le développement physiques, mentaux et émotionnels des enfants. Elles mettront en œuvre des mesures d'atténuation appropriées et surveilleront les risques après le déploiement, notamment en détectant les incidents graves impliquant des enfants et en y répondant.

Les organisations annonçant des contributions au titre de la présente Charte s'engagent à respecter ces principes et garanties dans le cadre de la Coalition. Elles s'engagent également à rendre compte de manière transparente des progrès accomplis au regard des résultats convenus et de la mise en œuvre de ces principes. Les présidents du Comité scientifique publieront chaque année un bilan des progrès réalisés dans l'ensemble du programme de recherche, présentant notamment les productions, les lacunes dans les connaissances et les priorités de recherche actualisées. Afin de favoriser la transparence et d'éviter la multiplication des obligations de déclaration, les signataires sont encouragés à rendre compte chaque année de leurs progrès au moyen des mécanismes de suivi existants.

Annexe 1 : Organisation du programme de recherche

Le programme de recherche indépendant s'articule autour de trois axes de travail complémentaires, présentés ci-dessous. Les priorités scientifiques, les méthodes et les résultats demeurent sous responsabilité scientifique indépendante.

A. Observatoire de l'humain et de l'IA : L'Observatoire définira d'abord les effets prioritaires sur le développement et les comportements des modèles à mesurer, sous réserve de l'approbation du Comité scientifique. Il suivra ensuite les comportements des systèmes d'IA ainsi que les usages et les expériences des enfants selon les âges, les pays et les contextes culturels, afin de repérer les tendances émergentes et les lacunes de la recherche.

B. Mesures communes : Il s'agit d'un axe de recherche auquel les chercheurs et les équipes techniques des entreprises participeront aux côtés de chercheurs indépendants. Ils élaboreront, testeront et valideront des mesures communes et des outils open source, notamment des classificateurs et des systèmes de notation, afin de rendre opérationnelles les priorités définies par l'Observatoire. Ces outils permettront de mener des recherches comparables entre les produits, les pays et dans le temps. Les entreprises pourront les appliquer dans leurs systèmes et partager des résultats codés, limitant ainsi la nécessité de transférer des conversations complètes. Des financements industriels mutualisés et des contributions en nature soutiendront ces travaux, sous supervision scientifique indépendante.

C. Données probantes et effets sur le développement : Des revues de la littérature scientifique, des expérimentations et des études longitudinales examineront les effets de l'IA sur le développement et le bien-être des enfants selon les âges et les contextes. Les chercheurs pourront utiliser les mesures communes et les résultats de l'Observatoire tout

en restant libres de développer leurs propres méthodes et de collecter des données de manière indépendante. Les résultats éclaireront la conception des produits, les politiques publiques et les travaux distincts de la Coalition sur l'évaluation et les normes.

Annexe 2 : Méthodes d'évaluation et communication des résultats

Les méthodes de recherche et d'évaluation, ainsi que la communication des résultats, devraient :

1. **Définir clairement les risques et les critères d'évaluation : Préciser les préjudices concrets évalués, ce qui constitue un échec et la manière dont le test mesure ces préjudices. Fournir des exemples de situations constituant ou non un échec afin de garantir une notation cohérente. Distinguer les comportements préjudiciables d'une discussion appropriée de sujets sensibles.**
2. **Refléter les besoins de développement et la diversité des enfants : Préciser les âges, les stades de développement, les langues, les capacités et les contextes couverts par l'évaluation. Signaler les différences significatives entre les groupes et identifier les populations pour lesquelles les données demeurent limitées.**
3. **Tester les usages réalistes et les détournements raisonnablement prévisibles : Concevoir des tests reflétant la manière dont les enfants et les adolescents communiquent et utilisent réellement l'IA, notamment le langage propre à leur âge, les révélations indirectes, la curiosité, l'expérimentation et les tentatives de contournement des protections. Inclure les interactions ordinaires et les situations ambiguës, et examiner les interactions prolongées et répétées lorsque des risques peuvent apparaître avec le temps.**
4. **Utiliser des mesures valides et transparentes : Documenter la sélection des tests, les critères de notation, les qualifications et le degré d'accord des évaluateurs, ainsi que les systèmes et configurations testés. Communiquer les instructions soumises aux modèles, les éléments pertinents de l'historique des conversations et les réponses obtenues, sous réserve de protections appropriées pour les informations sensibles. Signaler les incertitudes, les faux positifs et les cas non détectés lorsque cela est pertinent, et fournir suffisamment de détails méthodologiques pour permettre un examen indépendant et, lorsque cela est possible, la reproduction des travaux.**
5. **Évaluer l'efficacité concrète des mesures de protection : Évaluer l'ensemble des dispositifs de sécurité et l'expérience utilisateur, notamment la sécurité du modèle de base, la vérification de l'âge, les protections au niveau du produit, les mesures au niveau du compte et les notifications aux parents, le cas échéant. Vérifier que les utilisateurs de moins de 18 ans sont correctement identifiés et bénéficient des protections appropriées. Tester si ces éléments fonctionnent ensemble pour assurer une protection rapide et efficace, et évaluer les conséquences involontaires, notamment les risques pour la vie privée ou les restrictions inutiles à des usages bénéfiques.**
6. **Coordonner la configuration appropriée des tests : Travailler avec les entreprises concernées pour s'assurer que les comptes de test destinés aux moins de 18 ans sont configurés de manière appropriée et reflètent l'expérience effectivement déployée, tout en préservant l'indépendance des évaluateurs dans la conception des tests, la notation et les conclusions. Distinguer clairement et présenter séparément les tests de comptes de mineurs correctement configurés de ceux portant sur un âge inconnu ou mal classé, ou sur le contournement des protections liées à l'âge.**

La présente Charte a reçu le soutien d'Anthropic, de Black Forest Labs, de Microsoft et d'OpenAI.