Centrum für Europäische Politik
FREIBURG | BERLIN

Adhoc du cep

N° 25 | 2024 28 novembre 2024

Vers une gouvernance robuste de l'IA en Europe

Recommandations techniques pour le code de pratique de l'IA à usage général

Anselm Küsters

Source : DALL-E via ChatGPT incitation

L'intelligence artificielle transforme les industries et la vie quotidienne, ce qui a incité l'UE à élaborer un code de pratique général sur l'IA pour guider son utilisation sûre et éthique. Cette analyse porte sur la première version du code et propose des améliorations techniques pour mieux soutenir l'innovation, traiter les risques systémiques et protéger les utilisateurs.

- Les principes d'alignement sur les valeurs de l'UE, de proportionnalité et de pérennité du code sont louables, mais doivent être mieux mis en œuvre. Afin de soutenir efficacement les PME et de favoriser l'innovation, le code devrait fournir des exemples concrets de parcours de conformité simplifiés et prendre en compte les défis spécifiques auxquels sont confrontés les fournisseurs de logiciels libres. En outre, le code doit explicitement donner la priorité à la liberté d'expression et à l'accès à l'information, étant donné que les modèles GPAI deviennent des plateformes fondamentales pour l'interaction sur l'internet.
- La taxonomie des risques systémiques devrait être élargie pour inclure les effets en cascade et les retombées, qui sont essentiels à l'identification complète des risques dans les systèmes d'IA interconnectés. Les méthodologies interdisciplinaires issues de l'analyse des systèmes complexes, de la théorie des réseaux et de l'économie comportementale peuvent enrichir les processus traditionnels d'évaluation des risques. Le code devrait également établir des niveaux de gravité clairs pour les risques et garantir une grande rigueur scientifique dans les évaluations, éventuellement en utilisant des mesures pour l'accord entre évaluateurs.
- L'évaluation continue des risques pendant le déploiement est essentielle et devrait inclure la collecte systématique de données sur les interactions entre l'homme et l'IA, tout en respectant la vie privée de l'utilisateur. L'accent mis par le Code sur l'utilisation de structures "si-alors" pour les mesures d'atténuation des risques et de solides indicateurs de performance clés pourrait permettre des auto-évaluations automatisées, mais il faut veiller à éviter une simplification excessive. Les exigences de transparence pour les fournisseurs en aval devraient être modulées en fonction de leurs capacités, tandis que les rapports sur la consommation d'énergie devraient être encouragés.

Contenu

1 Introduction : Le code de pratique de l'IA à usage général 3
2 Observations générales 4
2.1 Alignement sur les principes de l'Union et adaptation à l'avenir 4
2.2 Proportionnalité et soutien aux start-ups, aux PME et aux fournisseurs de logiciels libres 5
2.3 L'AMPI en tant que nouveau changement de plateforme et la liberté d'expression 5
3 Observations du groupe de travail 2 6
3.1 Taxonomie des risques systémiques 6
3.2 Inclusion de la "perte de contrôle" 7
3.3 Propensions dangereuses des modèles et collusion de l'IA 8
3.4 Cadre de sécurité et de sûreté (SSF) 8
3.5 Méthodologie d'analyse des risques 9
3.6 Niveaux de gravité 9
3.7 Collecte de preuves et évaluation 10
3.8 Cycle de vie de l'évaluation des risques et données sur les interactions avec les utilisateurs 10
4 Observations granulaires 11
4.1 Mesures conditionnelles d'atténuation des risques à l'aide de structures "si-alors 11
4.2 Robustesse face aux erreurs de spécification 11
4.3 Exigences de transparence pour les fournisseurs en aval 11
4.4 Transparence des données détaillées pour les développeurs 12
4.5 Rapports sur la consommation d'énergie 12
4.6 Limites de robots.txt en matière de propriété des données 13
4.7 Atténuation des risques par des messages-guides (obligatoires) du système 13
4.8 Formats innovants pour les rapports sur la sûreté et la sécurité 14
4.9 Automatiser les décisions de développement et de déploiement 15
4.10 Définition des incidents graves 15
4.11 Sanctions en cas d'absence de notification 16
4.12 Critères de classification dynamique au-delà des seuils de calcul 16
5 Conclusion 17

Chiffres

Fig. 1 : Exemple empirique d'utilisation de messages-guides pour améliorer le comportement du modèle 14

1 Introduction : Le code de pratique de l'IA à usage général

En novembre 2024, l'approche proactive de l'Union européenne (UE) en matière de réglementation de l'intelligence artificielle (IA) a franchi une nouvelle étape avec la publication du premier projet de code de bonnes pratiques sur l'IA à usage général.[1] Ce projet, préparé par des experts indépendants et facilité par l'Office européen de l'IA, vise à s'aligner sur la future loi européenne sur l'IA et à la soutenir. Cette dernière définit un modèle d'IA à usage général (GPAI) comme tout modèle d'IA "qui fait preuve d'une grande généralité et est capable d'exécuter avec compétence un large éventail de tâches distinctes, quelle que soit la manière dont le modèle est mis sur le marché, et qui peut être intégré dans divers systèmes ou applications en aval, à l'exception des modèles d'IA utilisés pour des activités de recherche, de développement ou de prototypage avant d'être mis sur le marché".[2] Ces modèles, qui sont généralement de grands modèles de langage (LLM) comme le GPT-4 ou des modèles de génération d'images comme le DALL-E, se caractérisent donc par leurs capacités importantes et leur aptitude à être intégrés dans diverses applications en aval. À cet égard, la loi sur l'IA adopte une approche réglementaire fondée sur le risque. Alors que tous les fournisseurs de GPAI sont soumis à des obligations de base, ceux qui sont considérés comme présentant un "risque systémique" (présumé pour les modèles entraînés avec plus de 10^25 opérations en virgule flottante) sont soumis à des exigences supplémentaires. Les modèles GPAI devenant de plus en plus importants dans la vie quotidienne, ils nécessitent un cadre solide pour garantir leur utilisation en toute sécurité.

Le processus actuel de consultation des parties prenantes sur le "code de bonnes pratiques" susmentionné vise à harmoniser les efforts des développeurs, des régulateurs et des utilisateurs pour faire face aux nouveaux risques posés par les modèles GPAI, tout en contribuant à la sécurité juridique et en réduisant les charges bureaucratiques. Avec près de 1 000 participants, la consultation vise à recueillir des points de vue et des expertises variés, ce qui est essentiel compte tenu de l'évolution rapide des technologies de l'IA. Le Centre for European Policy (cep) contribue activement au groupe de travail 2, en se concentrant sur la taxonomie des risques et sur la taille et la capacité des fournisseurs. Cet engagement est fondé sur la conviction qu'une réglementation efficace de l'IA doit équilibrer l'innovation et les garde-fous afin de favoriser un environnement dans lequel l'IA peut se développer de manière responsable. Les politiques doivent refléter à la fois les nuances technologiques de l'IA et les réalités socio-économiques des parties prenantes. En particulier, pour réglementer efficacement l'IA, il faut reconnaître ses caractéristiques uniques, à savoir une croissance exponentielle, un développement itératif autonome et des processus internes souvent opaques (nature de "boîte noire").[3] Les modèles réglementaires traditionnels pourraient ne pas suffire ; nous préconisons plutôt une approche itérative et ascendante qui favorise la maîtrise de l'IA et minimise les risques systémiques. Une telle approche doit être adaptative et à l'épreuve du temps.

Cette note politique résume notre première série de commentaires sur le premier projet de code de pratique sur l'IA à usage général (le "code"). La section 2 présente des observations générales sur les principes généraux qui façonnent le code, en soulignant son alignement sur les valeurs de l'UE, sa proportionnalité et son caractère évolutif. La section 3 examine les dispositions spécifiques relatives au groupe de travail 2, en se concentrant sur l'identification et l'évaluation des risques systémiques, et analyse la taxonomie des risques systémiques et les méthodes d'évaluation des risques. La section 4 présente des observations plus détaillées qui vont au-delà de l'objectif du groupe de travail 2, en abordant diverses "mesures" énumérées dans le code et en suggérant des améliorations sur la base du projet accessible au public. Il est important de noter que cette contribution se limite à des observations générales sur le premier projet accessible au public. En tant que membre du groupe de travail,

---

[1] Le projet est accessible ici : https://digital-strategy.ec.europa.eu/en/library/first-draft-general-purpose-ai-code-practice-published-written-independent-experts (en bref : "Code de conduite, 1er projet").

[2] Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées en matière d'intelligence artificielle et modifiant les règlements (CE) n° 300/2008, (UE) n° 167/2013, (UE) n° 168/2013, (UE) 2018/858, (UE) 2018/1139 et (UE) 2019/2144 ainsi que les directives 2014/90/UE, (UE) 2016/797 et (UE) 2020/1828, PE/24/2024/REV/1, JO L, 2024/1689, 12.7.2024 (en bref : "Loi sur l'IA de l'UE"), art. 3, para. 63.

[3] Voir, par exemple, Küsters et Vöpel (2023), https://commongroundeurope.eu/blog/vorfahrtsregeln-stattstoppschilder-warum-europa-denanschluss-bei-kuenstlicherintelligenz-gerade-jetzt-nichtverlieren-darf/.

l'auteur, ainsi que les autres membres universitaires, ont accepté de préserver la confidentialité des discussions internes pendant les réunions.

2 Observations générales

2.1 Alignement sur les principes de l'Union et adaptation à l'avenir

Le code de pratique s'appuie sur six considérations clés, énoncées au début du document :

- Alignement sur les principes et les valeurs de l'Union
- Alignement sur la loi sur l'IA et les approches internationales
- Proportionnalité aux risques
- Proportionnalité à la taille et aux capacités des fournisseurs
- Soutien et croissance de l'écosystème de la sécurité de l'IA
- La protection de l'avenir

Dans cette liste, l'accent mis sur la protection de l'avenir est particulièrement important. Compte tenu du développement exponentiel et de la nature "boîte noire" de l'IA, les cadres réglementaires doivent être dynamiques et adaptables.[4] Le projet suggère à juste titre qu'"un équilibre devrait être trouvé entre des exigences concrètes et une flexibilité permettant d'adapter et de mettre à jour les règles au fur et à mesure de l'évolution de la technologie et de l'industrie".[5] Cette approche est conforme aux perspectives académiques sur la gouvernance agile de l'IA, qui préconisent des mécanismes réglementaires capables d'évoluer avec les progrès technologiques.[6] Pour rendre opérationnelle la protection de l'avenir, le code pourrait comporter, au-delà des cycles de révision réguliers promis, des mécanismes tels que des clauses d'extinction pour des règles spécifiques ou des voies de conformité adaptatives qui tiennent compte de l'émergence de nouvelles technologies. Comme indiqué dans le projet de code lui-même, des références à des sources dynamiques telles que des bases de données d'incidents et des normes consensuelles pourraient contribuer à garantir que les fournisseurs sont tenus au courant des derniers développements et des meilleures pratiques.

Le premier projet mis à disposition à la mi-novembre reconnaît qu'il "n'inclut pas encore de section sur la manière dont le Code sera révisé et mis à jour - cela sera inclus dans les versions ultérieures du projet de Code".[7] L'objectif est de mettre en place un mécanisme de révision et de mise à jour permanentes afin de garantir que le code reste pertinent et efficace. À cet égard, un processus de révision transparent et accessible est essentiel au succès à long terme du code. La création d'un portail dédié à la contribution de la société civile, où les parties prenantes peuvent facilement soumettre leurs commentaires et partager des exemples (tels que des captures d'écran de problèmes avec les interfaces GPAI comme les chatbots), encouragerait une participation inclusive. Des itérations fréquentes sont essentielles ; attendre trop longtemps entre les mises à jour pourrait rendre le code obsolète en raison du rythme rapide du développement de l'IA. La transparence du processus d'examen renforce la confiance et la responsabilité et encourage ainsi l'engagement actif des parties prenantes. Cette approche est cohérente avec les modèles de gouvernance adaptative, qui mettent l'accent sur l'élaboration itérative de politiques dans des environnements technologiques dynamiques.[8]

---

4 Küsters (2024), https://www.cep.eu/eu-topics/details/anticipating-ai-instead-of-preventing-it.html.

5 Code de conduite, 1er projet, p. 4.

6 Pour une vue d'ensemble, voir : Araz Taeihagh, "Governance of Artificial Intelligence", *Policy and Society* 40, no. 2 (3 avril 2021) : 137–57, https://doi.org/10.1080/14494035.2021.1928377.

7 Code de conduite, 1er projet, p. 3.

8 Pour le cas des véhicules autonomes, voir : Araz Taeihagh et Hazel Si Min Lim, "Governing Autonomous Vehicles : Emerging Responses for Safety, Liability, Privacy, Cybersecurity, and Industry Risks', *Transport Reviews* 39, no 1 (2 janvier 2019) : 103–28, https://doi.org/10.1080/01441647.2018.1494640.

2.2 Proportionnalité et soutien aux start-ups, aux PME et aux fournisseurs de logiciels libres

Le projet reconnaît de manière appropriée la diversité du paysage des fournisseurs d'IA, en soulignant que "les obligations applicables aux fournisseurs de modèles d'IA à usage général devraient tenir dûment compte de la taille du fournisseur de modèles d'IA à usage général et permettre des parcours de conformité simplifiés pour les PME et les start-ups disposant de moins de ressources financières que celles qui sont à l'avant-garde du développement de l'IA".[9] Cette différenciation est en effet cruciale pour favoriser l'innovation à long terme. Toutefois, pour atteindre pleinement cet objectif, le code devrait fournir des exemples concrets de la manière dont les mesures peuvent être simplifiées pour les PME. Par exemple, l'offre de modèles de conformité standardisés, d'exemptions concrètes de certaines obligations de déclaration, d'un accès subventionné à des cadres ou installations de test tiers et de programmes d'assistance technique pour la mise à jour des modèles d'IA pourrait réduire les obstacles à la conformité. Alors que les jeunes entreprises se lancent dans des innovations plus risquées,[10] un soutien réglementaire adapté peut améliorer de manière significative leur capacité à innover de manière responsable.

Dans le même ordre d'idées, le Code stipule que "les mesures, sous-mesures et ICP [indicateurs clés de performance] doivent être proportionnées aux risques, c'est-à-dire qu'elles doivent être (a) appropriées pour atteindre l'objectif visé, (b) nécessaires pour atteindre l'objectif visé, et (c) ne pas imposer une charge excessive par rapport à l'objectif visé".[11] Il s'agit d'un principe sain, qui reconnaît la nécessité de trouver un équilibre entre la réduction des risques et les réalités pratiques de la mise en œuvre. Les parties prenantes de l'industrie apprécieront probablement cette approche qui réduit les coûts de mise en conformité inutiles et encourage l'innovation. Toutefois, l'impact réel dépendra de la manière dont le principe sera mis en œuvre. Reconnaissant "l'impact positif que les modèles open-source ont eu sur le développement de l'écosystème de la sécurité de l'IA",[12] le projet envisage judicieusement des exemptions et des obligations adaptées pour les fournisseurs open-source. Comme l'ont noté les experts, l'IA à code source ouvert joue un rôle essentiel dans la démocratisation de l'accès aux technologies de l'IA sans accroître les risques à la marge.[13] Néanmoins, le code devrait définir des seuils à partir desquels les modèles open-source sont soumis à certaines réglementations, peut-être en fonction de l'ampleur de leur déploiement ou de leur potentiel d'impact "grave".

Dans l'ensemble, une question clé à l'avenir sera de savoir s'il convient de soutenir les PME et les fournisseurs de logiciels libres a) en les exemptant de certaines règles (par exemple, les lourdes obligations de transparence) ou b) en facilitant simplement leur conformité aux règles, qui sont les mêmes pour chaque modèle de l'AMPI. Du point de vue de la compétitivité, l'option a) pourrait être préférable, tandis qu'une attention stricte à la sécurité de l'IA soutiendrait l'option b), étant donné que les modèles GPAI des petits fournisseurs pourraient encore présenter des risques systématiques. D'après le projet actuel, il semble que l'Office AI penche davantage vers la deuxième approche (option b).

2.3 L'AMPI en tant que nouvelle plateforme et la liberté d'expression

Les LLM et autres systèmes GPAI apparaissent rapidement comme des plateformes fondamentales pour l'accès à l'information et l'interaction sur l'internet, marquant un changement de paradigme important dans la manière dont nous consommons et traitons l'information. Si cette évolution est prometteuse à bien des égards, elle soulève également des préoccupations majeures quant à la préservation de la liberté d'expression et de l'accès équitable à l'information. Comme l'affirme Mme Calvet-Bademunt, il est urgent de se préoccuper du risque que les systèmes d'intelligence artificielle censurent ou biaisent l'information par inadvertance, une préoccupation qui

[9] Code de conduite, 1er projet, p. 4.
[10] Alex Coad, Agustí Segarra, et Mercedes Teruel, 'Innovation and Firm Growth : Does Firm Age Play a Role ?", *Research Policy* 45, no. 2 (mars 2016) : 387–400, https://doi.org/10.1016/j.respol.2015.10.015.
[11] Code de conduite, 1er projet, p. 3.
[12] Code de conduite, 1er projet, p. 4.
[13] Sayash Kapoor et al, 'On the Societal Impact of Open Foundation Models' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2403.07918.

deviendra de plus en plus pressante au fur et à mesure que ces technologies s'intègreront dans notre vie quotidienne.[14] Dans ce contexte, l'élaboration du code de pratique GPAI de l'UE est une occasion cruciale de sauvegarder la liberté d'expression dans le cadre mondial de la gouvernance de l'IA. Comme le conclut Mme Calvet-Bademunt, le code doit explicitement aborder les risques systémiques associés à la modération du contenu, au filtrage de l'information et à la possibilité pour l'IA d'influencer de manière significative le discours public. En particulier, il doit aller au-delà des vagues références aux "droits fondamentaux" et donner explicitement la priorité à la liberté d'expression. S'inspirant des normes relatives aux droits de l'homme, le code devrait exiger des fournisseurs d'IA qu'ils élaborent des politiques d'utilisation claires, détaillées et accessibles au public. En outre, toute restriction de contenu mise en œuvre par les systèmes d'IA devrait respecter les principes de légitimité et de proportionnalité, en veillant à ce que les restrictions d'expression soient bien justifiées et ne soient pas trop larges. Les stratégies d'atténuation des risques techniques devraient privilégier des méthodes telles que le prédémontage, le démantèlement et le contre-discours plutôt que la censure pure et simple lorsqu'il s'agit d'aborder des questions telles que la désinformation ou les préjugés. Cette approche préserverait la capacité des chercheurs et du grand public à explorer un large éventail d'idées à l'aide d'outils d'IA.

3 Observations du groupe de travail 2

3.1 Taxonomie des risques systémiques

La mesure 6 introduit une taxonomie pour identifier et catégoriser les risques systémiques associés aux modèles d'IA à usage général. Elle vise à fournir un cadre structuré permettant aux fournisseurs d'identifier les risques potentiels susceptibles d'avoir un impact généralisé. À l'heure actuelle, cette taxonomie répertorie les risques liés aux éléments suivants les cybercapacités offensives (telles que la découverte ou l'exploitation de vulnérabilités) ; la science à double usage permettant des attaques chimiques, biologiques, radiologiques et nucléaires par le développement, la conception, l'acquisition ou l'utilisation d'armes ; les problèmes découlant de l'incapacité à contrôler de puissants modèles d'IA autonomes ("perte de contrôle") ; le potentiel de développement accéléré et imprévisible de l'IA par l'utilisation automatisée de modèles pour la recherche et le développement de l'IA ; la facilitation de la persuasion et de la manipulation à grande échelle, y compris la désinformation ou la mésinformation qui menace les valeurs démocratiques et les droits de l'homme ; et la discrimination illégale à grande échelle à l'encontre des individus, des communautés ou des sociétés.[15]

Toutefois, la taxonomie actuelle décrite dans la mesure 6.1 **n'inclut pas explicitement les effets en cascade ou de débordement**, qui sont essentiels en période d'évolution technologique et sociétale rapide. Ces effets font référence à des situations où une défaillance ou une conséquence involontaire dans une partie du système déclenche une réaction en chaîne qui conduit à des problèmes systémiques plus vastes. Nos recherches antérieures au sein du CEP ont montré que les modèles d'IA formés sur des données historiques peuvent ne pas fonctionner de manière fiable lors d'événements sans précédent dans une "polycrise", affectant potentiellement d'autres parties critiques d'un système - un phénomène observé lors de la pandémie COVID-19, lorsque les modèles prédictifs n'ont pas réussi à prendre en compte les changements soudains dans le comportement des consommateurs et la dynamique du marché, avec des conséquences allant de l'évaluation automatisée du crédit au déploiement des forces de police.[16] L'absence de risques en cascade dans la taxonomie peut conduire à une sous-estimation du potentiel des modèles GPAI à contribuer à des défaillances systémiques, en particulier dans les systèmes interconnectés tels que la finance, les soins de santé ou d'autres infrastructures critiques. Cette omission est surprenante étant donné qu'une section ultérieure souligne que, outre les capacités inhérentes des modèles d'IA, les contextes sociotechniques dans lesquels ils opèrent devraient également être pris en compte lors de l'évaluation des risques systémiques (mesure 6.3.3).

---

[14] Ici et dans le reste de cette section, je m'appuie sur : Sauvegarder la liberté d'expression à l'ère de l'IA | TechPolicy.Press.

[15] Code de conduite, 1er projet, p. 17.

[16] Anselm Küsters, "AI as Systemic Risk in a Polycrisis", cepAdhoc (Berlin : Centre for European Policy, 13 décembre 2022), https://www.cep.eu/eu-topics/details/ai-as-systemic-risk-in-a-polycrisis-cepadhoc.html.

Cette approche holistique est soutenue par la théorie des systèmes socio-techniques, qui préconise l'analyse des systèmes technologiques dans leurs contextes sociaux plus larges.[17] L'inclusion des effets en cascade dans la taxonomie des risques systémiques est également justifiée par la théorie de la complexité, qui a récemment été appliquée à la sécurité de l'IA.[18] Comme indiqué ci-dessus, les effets en cascade représentent une catégorie unique de risque qui peut s'amplifier et se propager dans différents secteurs, ce qui peut avoir des conséquences imprévues et profondes. Cet aspect est crucial dans le contexte de systèmes d'IA complexes qui sont de plus en plus interconnectés (par exemple, par le biais d'API). En ce sens, le code devrait s'éloigner d'une approche réductionniste qui cherche à comprendre les systèmes d'IA et leurs risques en les décomposant, et utiliser plutôt d'autres méthodes (voir la section 3.5 ci-dessous) pour analyser les systèmes d'IA complexes constitués de nombreuses parties indépendantes qui interagissent pour produire des effets plus importants que leur somme. En intégrant les effets en cascade dans la taxonomie, le code fournirait un cadre plus solide pour l'évaluation des risques face au déploiement rapide des technologies et aux liens mondiaux.

En réponse à la question de savoir quels sont les considérations ou critères pertinents à prendre en compte pour déterminer si un risque est systémique (page 18 du projet de code), l'inclusion des effets en cascade justifie les critères suivants :

- Interconnectivité : La mesure dans laquelle le système d'IA interagit avec d'autres systèmes.
- Complexité : La possibilité que de petits changements aient des effets significatifs en raison de la dynamique du système ou d'une relation non linéaire entre la cause et l'effet.
- Sensibilité au temps : La vitesse à laquelle les effets peuvent se propager dans les systèmes.

En résumé, les effets en cascade et les retombées devraient être inclus dans la taxonomie des risques systématiques afin de traiter les risques directs et indirects du déploiement à grande échelle de l'IA, en réorientant l'attention des développeurs et des politiciens du GPAI vers des stratégies d'atténuation plus efficaces à l'échelle de la société.

Plus généralement, plusieurs autres risques théoriques pourraient être ajoutés à la taxonomie. Il s'agit notamment des risques environnementaux liés à l'adoption généralisée de l'IA ou, comme le montrent nos recherches au sein du CEP, des risques sociétaux liés à l'augmentation du chômage si les modèles d'IA ont des effets substitutifs plutôt que complémentaires sur le marché du travail.[19] Cela soulève la question de savoir pourquoi cette liste particulière de risques a été choisie par l'Office de l'IA. Une méthode plus structurée et plus transparente de création de la taxonomie des risques serait donc la bienvenue. L'un des moyens d'y parvenir consisterait à lier plus étroitement les risques identifiés à la formulation de la loi initiale sur l'AI.

3.2 Inclusion de la "perte de contrôle"

Dans la mesure 6.1, le Code identifie la "perte de contrôle" comme un risque systémique,[20] faisant probablement référence à des scénarios dans lesquels les fournisseurs sont incapables de contrôler de puissants modèles d'IA autonomes.

S'il existe un potentiel théorique pour que les modèles d'IA deviennent incontrôlables, la formulation de ce risque doit être examinée avec soin. La notion de systèmes d'IA menant à une "perte de contrôle" découle souvent de récits spéculatifs où l'intelligence artificielle générale (AGI) surpasse les capacités humaines, comme l'explique Bostrom.[21] Toutefois, des spécialistes de l'IA critique comme Whittaker ont fait remarquer à plusieurs reprises

---

[17] Gordon Baxter et Ian Sommerville, "Socio-Technical Systems : From Design Methods to Systems Engineering', *Interacting with Computers* 23, no. 1 (janvier 2011) : 4–17, https://doi.org/10.1016/j.intcom.2010.07.003.

[18] Dan Hendrycks (à paraître), Introduction to AI Safety, Ethics, and Society, chapitre 5, disponible ici : 5.1 : Systèmes complexes - Manuel sur la sécurité, l'éthique et la société de l'IA.

[19] Küsters et Poli (2024), Résister ou se relancer face à la montée des robots (cepStudie) | cep - Centrum für europäische Politik.

[20] Code de conduite, 1er projet, p. 17.

[21] Nick Bostrom, *Superintelligence : Paths, Dangers, Strategies*, réimpression avec corrections 2017 (Oxford, Royaume-Uni : Oxford University Press, 2017).

que le fait de se concentrer sur ces risques spéculatifs peut détourner l'attention de problèmes immédiats et vérifiables tels que les biais algorithmiques, les violations de la vie privée et les vulnérabilités en matière de sécurité.[22] Une taxonomie plus équilibrée omettrait peut-être cette question pour se concentrer sur l'explicabilité, qui peut atténuer les risques associés aux comportements involontaires sans invoquer de scénarios spéculatifs.[23]

3.3 Propensions dangereuses des modèles et collusion de l'IA

La mesure 6.3.2, qui énumère les propensions dangereuses des modèles, aborde explicitement la possibilité que les modèles d'IA s'engagent dans une collusion avec d'autres systèmes d'IA.[24]

L'idée que les modèles d'IA puissent être "collusoires" a été explorée dans le contexte du droit de la concurrence. Des chercheurs tels qu'Ezrachi et Stucke ont théorisé le fait que l'IA pourrait permettre une collusion tacite, où les algorithmes apprennent indépendamment à coordonner les stratégies de prix sans communication explicite.[25] Bien que cet argument conceptuel ait été mis en avant pendant un certain nombre d'années, les preuves empiriques d'une collusion induite par l'IA ont été limitées. Des études ultérieures suggèrent que si l'IA peut faciliter la collusion dans certaines conditions, la dynamique du marché ([26]) et la diversité des algorithmes utilisés par les différentes entreprises rendent improbable une collusion généralisée. Comme le note Dorner, "la littérature manque souvent de la perspective des informaticiens et semble régulièrement surestimer l'applicabilité des progrès récents en matière d'apprentissage automatique au problème complexe de coordination auquel les entreprises sont confrontées lorsqu'elles forment des cartels".[27] Par conséquent, s'il est prudent de considérer la collusion comme un risque potentiel, le code devrait fonder ses mandats sur des évaluations basées sur des preuves afin d'éviter d'accorder trop d'importance aux risques spéculatifs.

3.4 Cadre de sécurité et de sûreté (SSF)

La mesure 7 souligne que les fournisseurs de modèles d'IA à usage général doivent mettre en œuvre un cadre de sûreté et de sécurité (SSF) complet. Ce cadre est conçu pour évaluer et atténuer les risques systémiques associés aux modèles d'IA, en particulier ceux qui ont des capacités à fort impact. Il reconnaît que des mesures moins complètes peuvent être suffisantes lorsqu'un nouveau modèle possède des capacités similaires à celles déjà déployées en toute sécurité, sans que des risques importants ne se matérialisent.[28]

Cette dernière disposition est particulièrement bénéfique pour les chercheurs et les PME, qui doivent souvent rattraper les grands laboratoires d'IA bien financés et gérés par de grandes entreprises technologiques. Le fossé qui se creuse entre les leaders de l'industrie et les institutions universitaires ou plus petites a été documenté à maintes reprises par des mesures telles que les articles les plus cités, les brevets et les contributions aux principales conférences sur l'IA.[29] Les PME et les établissements universitaires ne disposent généralement pas des vastes ressources (informatiques et humaines) nécessaires aux contrôles de robustesse les plus rigoureux effectués par les développeurs d'IA de pointe. En autorisant la proportionnalité dans les mesures de conformité, le code reconnaît les contraintes de ressources auxquelles ces entités sont confrontées. Toutefois, il est important

---

[22] Voir : https://thebulletin.org/2024/07/three-key-misconceptions-in-the-debate-about-ai-and-existential-risk/.

[23] Finale Doshi-Velez et Been Kim, " Towards A Rigorous Science of Interpretable Machine Learning " (arXiv, 2017), https://doi.org/10.48550/ARXIV.1702.08608.

[24] Code de conduite, 1er projet, p. 21.

[25] Ariel Ezrachi et Maurice E. Stucke, *Virtual Competition : The Promise and Perils of the Algorithm-Driven Economy* (Cambridge, Massachusetts : Harvard University Press, 2016).

[26] Stephanie Assad et al, "Algorithmic Pricing and Competition : Empirical Evidence from the German Retail Gasoline Market", *SSRN Electronic Journal*, 2020, https://doi.org/10.2139/ssrn.3682021.

[27] Florian E. Dorner, "Algorithmic Collusion : A Critical Review" (arXiv, 2021), https://doi.org/10.48550/ARXIV.2110.04740.

[28] Code de conduite, 1er projet, p. 21.

[29] Nur Ahmed et Muntasir Wahed, "The De-Democratization of AI : Deep Learning and the Compute Divide in Artificial Intelligence Research" (arXiv, 2020), https://doi.org/10.48550/ARXIV.2010.15581.

de suivre cette approche dans le temps, car le paysage de l'industrie évolue et les PME développent potentiellement des modèles plus avancés qui peuvent encore justifier une surveillance plus stricte.

3.5 Méthodologie d'analyse des risques

Pour traiter efficacement les risques potentiels associés aux systèmes d'IA, les prestataires doivent utiliser des méthodologies d'analyse des risques solides (sous-mesure 9.1). Ces approches doivent non seulement identifier les voies potentielles par lesquelles leurs modèles d'IA pourraient créer des risques systémiques, mais aussi évaluer la probabilité et l'impact de ces risques.[30] Les recherches récentes sur la sécurité de l'IA ont commencé à considérer l'IA comme un système complexe, en explorant les comportements inattendus que ces systèmes présentent souvent.[31] Cette perspective remet en question le paradigme dit "réductionniste", qui suppose que les systèmes peuvent être compris en les décomposant en leurs éléments constitutifs, puisque les systèmes d'IA sont constitués de nombreuses parties indépendantes qui interagissent pour produire des effets supérieurs à leur somme. Par conséquent, les effets de l'IA sur la société, qui est elle-même un système complexe, sont intrinsèquement difficiles à prévoir et nécessitent des approches analytiques plus sophistiquées.

Les techniques traditionnelles d'analyse des risques auxquelles se réfère le projet de code, bien que précieuses, peuvent donc ne pas rendre compte de toute la complexité des systèmes d'IA et de leurs interactions avec les utilisateurs humains et les structures sociétales. Il est donc essentiel d'intégrer des méthodologies issues d'autres disciplines pour enrichir le processus d'évaluation. L'analyse des systèmes complexes, telle que proposée par Holland pour les "systèmes adaptatifs complexes", c'est-à-dire les systèmes comprenant de nombreux composants qui s'adaptent ou apprennent au fur et à mesure de leur interaction,[32] peut fournir des indications sur le comportement émergent des systèmes d'IA et aider à anticiper les conséquences imprévues qui peuvent résulter de l'interaction détaillée des composants du système. De même, la théorie des réseaux pourrait fournir des outils pour analyser la manière dont les systèmes interconnectés peuvent propager les risques, ce qui est particulièrement pertinent dans le contexte de "polycrise" mentionné plus haut.[33] En outre, les enseignements de l'économie comportementale, dont les pionniers sont des chercheurs tels que Tversky et Kahneman,[34] , peuvent aider à anticiper la manière dont les utilisateurs peuvent interagir avec les modèles GPAI de manière involontaire en raison de biais cognitifs, ce qui pourrait exacerber ou créer des risques entièrement nouveaux.

3.6 Niveaux de gravité

Conformément à la mesure 9.3, les prestataires de GPAI classent les risques recensés par niveau de gravité, avec au moins un niveau où les risques sont inacceptables en l'absence de mesures de sauvegarde.[35] La création de ces niveaux de gravité facilite l'établissement de priorités dans la gestion des risques.

Bien qu'il n'existe pas, du moins à la connaissance de l'auteur, de norme universellement acceptée pour ces niveaux dans l'IA, les fournisseurs peuvent s'inspirer de la pensée systémique ainsi que des matrices de risques établies utilisées dans les secteurs où la sécurité est essentielle.[36] L'élaboration d'un consensus sur le concept de "gravité" devrait aller au-delà de l'approche proposée consistant à multiplier un risque spécifique par sa probabilité et inclure des critères qualitatifs tels que :

- Impact sur les droits de l'homme
- l'échelle, c'est-à-dire le nombre de personnes ou de systèmes potentiellement affectés

---

30 Code de conduite, 1er projet, p. 22.

31 Dan Hendrycks (2024), Introduction to AI Safety, Ethics, and Society, chapitre 5, disponible en ligne.

32 John H. Holland, "Studying Complex Adaptive Systems", *Journal of Systems Science and Complexity* 19, no. 1 (mars 2006) : 1–8, https://doi.org/10.1007/s11424-006-0001-z.

33 Küsters, "AI as Systemic Risk in a Polycrisis" (L'IA en tant que risque systémique dans une polycrise).

34 Amos Tversky et Daniel Kahneman, "Judgment under Uncertainty : Heuristics and Biases : Biases in Judgments Reveal Some Heuristics of Thinking under Uncertainty", *Science* 185, no. 4157 (27 septembre 1974) : 1124–31, https://doi.org/10.1126/science.185.4157.1124.

35 Code de conduite, 1er projet, p. 22.

36 Nancy G. Leveson, *Engineering a Safer World : Systems Thinking Applied to Safety* (The MIT Press, 2012), https://doi.org/10.7551/mitpress/8179.001.0001.

- L'irréversibilité, c'est-à-dire la possibilité d'annuler les conséquences.

Toutefois, l'inquiétude du public, c'est-à-dire le niveau d'anxiété de la société associé à un risque particulier, ne devrait pas influencer la définition de la "gravité", car le discours public est actuellement fortement influencé par des récits dystopiques de science-fiction (voir également le point 3.2 ci-dessus).

3.7 Collecte et évaluation des preuves

La mesure 10 engage les signataires à mettre en place un processus continu de collecte d'informations sur le risque systémique, en utilisant une série de méthodes telles que la prévision, l'analyse comparative, l'analyse des risques et la simulation.[37]

Le choix de la méthode d'évaluation appropriée dépend essentiellement du contexte spécifique. Certains risques sont mieux évalués à l'aide de méthodes spécifiques. Par exemple, les vulnérabilités en matière de sécurité peuvent nécessiter des tests contradictoires, tandis que la partialité peut être évaluée à l'aide de critères d'équité.[38] Toutefois, la puissance de calcul et l'expertise peuvent limiter le choix des méthodes, en particulier pour les PME. Pour déterminer si une évaluation est complète, des mesures analogues au Kappa de Fleiss pourraient être utilisées pour évaluer l'accord entre évaluateurs dans les évaluations qualitatives. Le kappa de Fleiss est une mesure statistique qui permet d'évaluer l'accord entre plusieurs évaluateurs, ce qui peut être crucial lorsqu'il s'agit de systèmes d'IA complexes où les jugements subjectifs peuvent jouer un rôle dans l'évaluation des risques. En quantifiant le niveau d'accord entre différents évaluateurs, le Kappa de Fleiss ou des mesures similaires fournissent une métrique pour la fiabilité du processus d'évaluation lui-même, ajoutant une couche supplémentaire de rigueur à la méthodologie d'évaluation des risques.

3.8 Cycle de vie de l'évaluation des risques et données sur les interactions avec les utilisateurs

La mesure 11 exige des signataires qu'ils évaluent en permanence les risques tout au long du cycle de vie du modèle d'IA, en particulier lors de son déploiement. Il s'agit notamment de mettre à jour les évaluations des risques au moins tous les six mois ou lorsque des changements significatifs interviennent, en tenant compte des données issues de la surveillance du modèle.[39]

Il est essentiel de mettre l'accent sur l'évaluation pendant la phase de déploiement, car de nombreux risques découlent des interactions réelles entre les utilisateurs et les modèles GPAI. Les évaluations statiques ne peuvent pas saisir la nature dynamique de l'interaction entre les utilisateurs et les systèmes d'IA. De nombreuses recherches montrent que les interactions homme-machine peuvent avoir des effets inattendus, notamment une confiance excessive dans les suggestions de l'IA ou une interprétation erronée des résultats.[40] Il est essentiel de comprendre cette dynamique pour atténuer les risques à long terme. Pour recueillir efficacement des données sur les interactions entre l'homme et l'IA, les fournisseurs de modes de transport pourraient mettre en place des canaux permettant aux utilisateurs de signaler des problèmes tels que des interprétations erronées ou des comportements involontaires des modèles, analyser les schémas d'interaction entre les utilisateurs et l'IA afin d'identifier les malentendus courants, ou expérimenter différentes versions de modèles afin d'évaluer les réponses des utilisateurs et les performances des modèles dans des conditions réelles (tests A/B). Toutefois, comme l'illustrent ces exemples, une telle collecte de données d'interaction avec l'utilisateur soulève également des questions en matière de protection de la vie privée. La conformité au règlement général sur la protection des données (RGPD)

---

37 Code de conduite, 1er projet, p. 23.

38 Ninareh Mehrabi et al, 'A Survey on Bias and Fairness in Machine Learning', *ACM Computing Surveys* 54, no. 6 (31 juillet 2022) : 1-35, https://doi.org/10.1145/3457607.

39 Code de conduite, 1er projet, p. 25.

40 Pour une vue d'ensemble, qui souligne le rôle de l'anthropomorphisme de l'IA, voir : Ella Glikson et Anita Williams Woolley, "Human Trust in Artificial Intelligence : Review of Empirical Research", *Academy of Management Annals* 14, no 2 (juillet 2020) : 627-60, https://doi.org/10.5465/annals.2018.0057.

est obligatoire. L'anonymisation des données, l'obtention d'un consentement éclairé et la garantie de la sécurité des données sont donc des conditions préalables importantes pour cette approche.

4 Observations granulaires

4.1 Mesures conditionnelles d'atténuation des risques à l'aide de structures "si-alors".

Le code propose de lier "les sous-mesures d'atténuation des risques aux IRC d'évaluation des risques, en utilisant par exemple des exigences de type "si-alors". Par exemple, si un modèle d'IA à usage général présentant un risque systémique est évalué comme ayant une capacité X, des mesures d'atténuation des risques Y doivent être mises en place, guidées par des ICP Z".[41]

Cette caractéristique promet d'améliorer la clarté des obligations de conformité envisagées et a des implications intéressantes du point de vue de la politique numérique. En structurant les exigences dans un format "si-alors", les fournisseurs peuvent non seulement mieux comprendre et mettre en œuvre les mesures nécessaires, mais ils pourraient même les automatiser, car cette logique est similaire aux structures de programmation. L'automatisation des contrôles de conformité pourrait grandement profiter aux PME et aux développeurs en leur fournissant des outils accessibles pour vérifier la conformité au code sans encourir des coûts juridiques ou d'expertise élevés. Toutefois, il convient de veiller à ce que les contrôles automatisés ne simplifient pas à l'excès les considérations éthiques inhérentes à l'utilisation des modèles GPAI.

4.2 Robustesse face aux erreurs de spécification

Le code insiste sur le fait que "les sous-mesures et les indicateurs clés de performance doivent également pouvoir être contournés ou mal spécifiés". Pour ce faire, le code peut, par exemple, éviter d'utiliser inutilement des termes ou des mesures de substitution. L'Office AI contrôlera et examinera les sous-mesures et les indicateurs de performance clés susceptibles d'être contournés ou de faire l'objet d'autres formes de spécification erronée".[42]

Il s'agit en effet d'une considération essentielle, car des paramètres ambigus ou mal définis peuvent conduire, intentionnellement ou non, à la non-conformité ou à des manquements à l'éthique. Par-dessus tout, pour garantir la robustesse, il faut utiliser un langage clair et sans ambiguïté et tenir compte des variations linguistiques et culturelles susceptibles d'affecter l'interprétation. La collaboration internationale dans le développement de l'IA signifie que les concepts et la terminologie peuvent être compris différemment selon les régions. Comme l'ont noté Waber et ses coauteurs : "Les réglementations relatives à l'IA et aux données connexes étant rarement uniformes d'une région à l'autre, il peut être difficile de s'y conformer. Pour résoudre ce problème, les entreprises doivent développer un modèle d'éthique de l'IA global et contextuel qui donne la priorité à la collaboration avec les équipes et les parties prenantes locales et qui délègue le pouvoir de décision à ces équipes locales."[43] Des examens réguliers par le bureau de l'IA peuvent aider à identifier et à corriger les domaines susceptibles d'être mal interprétés.

4.3 Exigences de transparence pour les fournisseurs en aval

Les fournisseurs en aval qui intègrent les modèles GPAI sont tenus de divulguer les détails techniques, notamment "une description de l'architecture du modèle, le type de modèle, la taille du contexte le cas échéant, le nombre total de paramètres du modèle et le nombre de paramètres actifs pendant l'inférence", ainsi que "les dépendances versionnées pour les logiciels et/ou le matériel requis".[44]

---

41 Code de conduite, 1er projet, p. 3.
42 Code de conduite, 1er projet, p. 4.
43 Voir : https://hbr.org/2024/08/how-companies-can-take-a-global-approach-to-ai-ethics.
44 Code de conduite, 1er projet, p. 11.

Si la transparence est essentielle (mais pas suffisante),[45] ces exigences peuvent être trop contraignantes pour les fournisseurs en aval, en particulier les PME ou les développeurs individuels. Dernièrement, l'intégration des modèles GPAI tels que les LLM a été grandement simplifiée (pensez à la série Llama de Meta accessible via GitHub) et les développeurs peuvent ne pas avoir une connaissance approfondie de l'architecture ou des paramètres du modèle sous-jacent, en particulier lorsqu'ils utilisent des modèles pré-entraînés ou des API. L'imposition d'exigences de divulgation aussi détaillées pourrait étouffer l'innovation et exclure les petits acteurs qui n'ont pas les ressources nécessaires pour s'y conformer. Une approche graduelle de la transparence pourrait être plus efficace, le niveau d'information requis variant en fonction des capacités concrètes des fournisseurs en aval.

4.4 Transparence des données détaillées pour les développeurs

Les développeurs sont tenus de divulguer "des informations spécifiques sur les données utilisées pour former/tester/valider le modèle, telles que la fraction des données provenant de différentes sources de données et les principales caractéristiques des données de formation, de test et de validation".[46]

Cette caractéristique du code actuel est tout à fait louable, car elle favorise la reproductibilité et la responsabilité dans le développement de l'IA. La transparence des sources de données et des caractéristiques de l'entraînement et des tests est essentielle pour comprendre les biais potentiels et les limites des modèles d'IA.[47] La communauté universitaire s'inquiète de plus en plus du manque de reproductibilité de la recherche sur l'apprentissage automatique en raison de l'insuffisance des rapports.[48] En particulier, le manque de séparation entre les données de test et les données d'entraînement a été critiqué.[49] En exigeant une transparence détaillée des données et de la formation, le code aligne les pratiques de l'industrie sur les normes universitaires, ce qui accroît la fiabilité des systèmes d'IA et favorise la confiance des utilisateurs.

4.5 Rapport sur la consommation d'énergie

Les promoteurs doivent "détailler les informations et la méthodologie qu'ils utilisent pour évaluer la consommation d'énergie [...], en cohérence avec tout acte délégué adopté conformément à l'article 97 de la loi sur l'énergie atomique pour détailler les méthodologies de mesure et de calcul en vue de permettre l'établissement d'une documentation comparable et vérifiable".[50]

Cette exigence répond aux préoccupations environnementales liées aux exigences informatiques de l'entraînement de grands modèles d'IA. Il s'agit là d'une bonne nouvelle, car nos recherches antérieures ont mis en évidence l'empreinte carbone significative de l'entraînement des modèles GPAI de pointe.[51] En exigeant la déclaration de la consommation d'énergie, le code encourage les développeurs à tenir compte de l'impact environnemental de leur travail, ce qui constitue un premier pas vers la réalisation des deux objectifs de l'UE en matière de transition. Il est approprié de limiter cette obligation aux développeurs initiaux, car ils maîtrisent le processus de formation et ont accès aux informations pertinentes. Les fournisseurs en aval utilisent généralement des modèles pré-entraînés et ont donc moins de contrôle sur la consommation d'énergie.

---

45 En ce qui concerne l'insuffisance de la transparence pour comprendre et gouverner les systèmes algorithmiques, voir : Mike Ananny et Kate Crawford, "Seeing without Knowing : Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability', *New Media & Society* 20, no. 3 (mars 2018) : 973–89, https://doi.org/10.1177/1461444816676645.

46 Code de conduite, 1er projet, p. 12.

47 Joy Buolamwini et Timnit Gebru, "Gender Shades : Intersectional Accuracy Disparities in Commercial Gender Classification', dans *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ed. Sorelle A. Friedler et Christo Wilson, vol. 81, Proceedings of Machine Learning Research (PMLR, 2018), 77-91, https://proceedings.mlr.press/v81/buolamwini18a.html.

48 Joelle Pineau et al, "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)", août 2021.

49 Justin Grimmer, Margaret E. Roberts et Brandon M. Stewart, *Text as Data : A New Framework for Machine Learning and the Social Sciences* (Princeton Oxford : Princeton University Press, 2022).

50 Code de conduite, 1er projet, p. 12.

51 Voir : https://www.cep.eu/eu-topics/details/environment-takes-a-backseat-in-eu-digital-push.html.

4.6 Limites de robots.txt en matière de propriété des données

Le code stipule que "[l]es ignatiers n'emploieront que des robots qui lisent et suivent les instructions exprimées conformément au protocole d'exclusion des robots (robots.txt)".[52]

Bien que le respect de robots.txt soit (du moins en théorie) une pratique courante, son efficacité dans le contexte de la collecte moderne de données d'IA est de plus en plus discutable. Robots.txt a été développé au tout début de l'internet et n'offre pas de protection juridique solide ni de mécanismes techniques d'application. Compte tenu des méthodes sophistiquées utilisées aujourd'hui pour la collecte de données, il n'est peut-être pas suffisant de s'appuyer uniquement sur le fichier robots.txt. Comme le note Pierce : "Pendant des décennies, robots.txt a régi le comportement des robots d'indexation. Mais comme les entreprises d'IA sans scrupules cherchent à obtenir de plus en plus de données, le contrat social de base du web est en train de s'effondrer".[53] Le Code devrait envisager d'imposer des protocoles plus avancés ou des accords juridiques pour respecter la propriété des données et la vie privée.

4.7 Atténuation des risques grâce à des messages-guides (obligatoires)

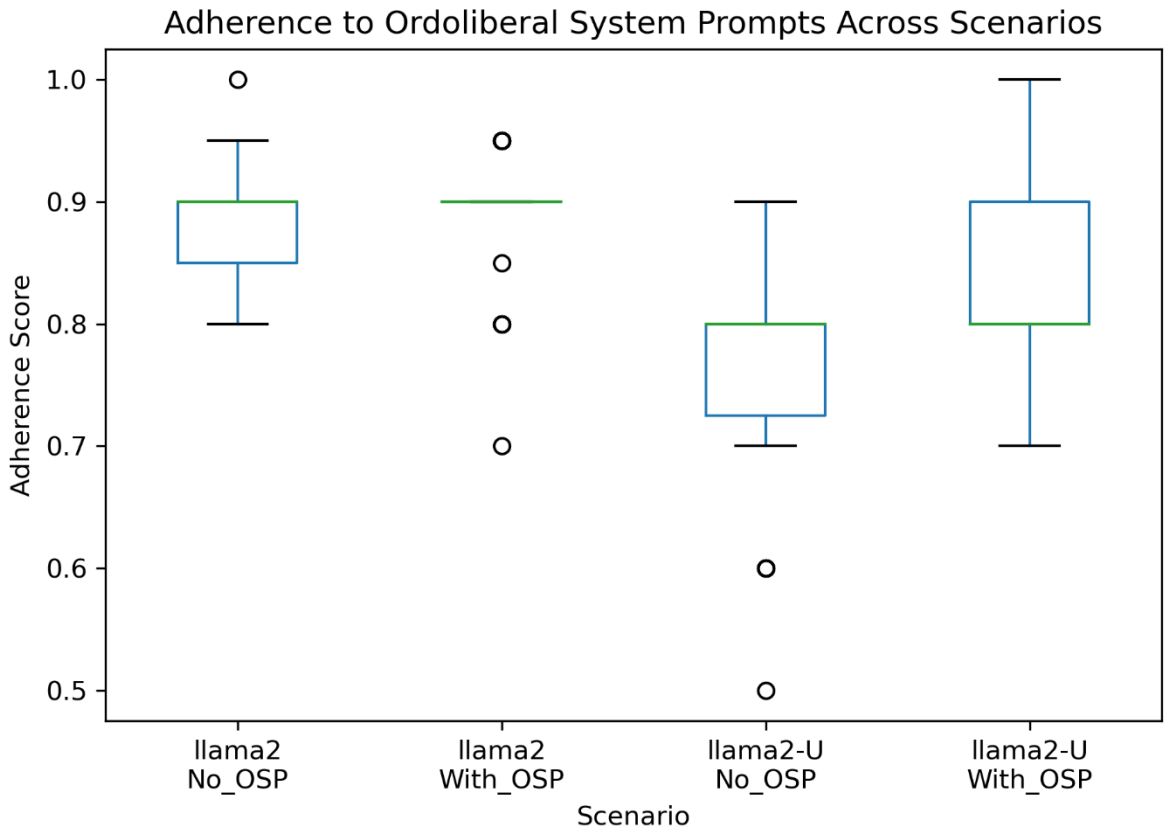
La mesure 12 suggère que les mesures d'atténuation de la sécurité pourraient inclure des "modifications comportementales d'un modèle", entre autres stratégies.[54] Mais comment y parvenir dans la pratique, c'est-à-dire au niveau technique ?

Comme nous l'avons indiqué dans nos recherches antérieures au sein du CEP, l'utilisation de ce que l'on appelle les "messages-guides du système", également connus sous le nom de "tuning" d'instructions, pourrait être un moyen rapide et largement efficace de guider les modèles d'IA vers des comportements sûrs et voulus.[55] Par exemple, l'incorporation d'instructions qui découragent la génération de contenus nuisibles ou de préjugés peut contribuer à atténuer certains risques. Par conséquent, le fait d'imposer certains types de messages-guides pourrait en partie normaliser les pratiques de sécurité dans l'ensemble de l'industrie. Toutefois, des exigences trop rigides pourraient étouffer la créativité ou avoir des conséquences inattendues, telles que la limitation des expressions légitimes de la liberté d'expression (voir le point 2.3 ci-dessus).

---

[52] Code de conduite, 1er projet, p. 15.
[53] Voir : https://www.theverge.com/24067997/robots-txt-ai-text-file-web-crawlers-spiders.
[54] Code de conduite, 1er projet, p. 27.
[55] Voir : https://www.cep.eu/eu-topics/details/in-search-of-laws-of-robotics.html.

Fig. 1: Exemple empirique d'utilisation de messages-guides pour améliorer le comportement du modèle



Source : Chiffre propre basé sur une recherche non publiée, à paraître sous le titre Küsters/Wörsdörfer, "Exploring Laws of Robotics".

Dans un article à paraître intitulé "Exploring Laws of Robotics : A Synthesis of Constitutional AI And Constitutional Economics", rédigé conjointement avec Manuel Wörsdörfer, nous avons étudié l'utilisation de messages-guides, en particulier de messages-guides ordolibéraux auto-conçus (OSP), comme méthode pratique pour guider les LLM vers un comportement éthique et intentionnel.[56] Nous avons mené une étude empirique comparant deux versions du modèle Llama 2 : le Llama 2 standard avec modération de contenu intégrée et le Llama 2-Uncensored, qui ne dispose pas de mécanismes de modération internes. En posant des questions éthiques aux deux modèles, avec et sans les OSP, nous avons constaté que les OSP avaient un effet négligeable sur le Llama 2 standard, mais qu'ils amélioraient de manière significative la conformité éthique du Llama 2-Uncensored. Cela suggère que si les messages-guides ne modifient pas de manière significative les modèles déjà équipés de couches de sécurité, ils peuvent améliorer efficacement le comportement des modèles dépourvus de tels mécanismes. Nos résultats suggèrent que les messages-guides constituent un outil facile à mettre en œuvre et à adapter pour améliorer le comportement éthique des systèmes d'IA, en particulier ceux qui ne sont pas censurés ou qui ne disposent pas de stratégies d'adaptation complètes, bien qu'ils ne remplacent pas des méthodes d'adaptation plus robustes avant l'entraînement.

4.8 Formats innovants pour les rapports de sûreté et de sécurité

Le code exige également des signataires qu'ils "créent un rapport de sûreté et de sécurité (SSR) pour tout modèle d'IA à usage général présentant un risque systémique".[57]

[56] Ce document fait actuellement l'objet d'une évaluation par les pairs. Une version antérieure, sans l 'expérience empirique mais avec les incitations du système ordolibéral, est accessible ici : https://www.cep.eu/eu-topics/details/in-search-of-laws-of-robotics.html.

[57] Code de conduite, 1er projet, p. 28.

Ces rapports sont généralement limités à des documents PDF, qui sont statiques et peuvent ne pas transmettre efficacement des informations complexes et techniques. Pour éviter de créer un autre document trop détaillé et difficile à comprendre, nous suggérons d'explorer d'autres formats plus innovants. L'exploration de formats alternatifs tels que les rapports computationnels utilisant les carnets Jupyter pourrait améliorer la transparence et la compréhension. Ces formats permettent une exploration interactive du code, des données et des résultats, ce qui facilite l'examen et la reproductibilité.[58] En particulier, l'inclusion d'éléments tels que des extraits de code, des visualisations de données et des cellules exécutables pourrait rendre les rapports de recherche scientifique plus informatifs et accessibles aux parties prenantes ayant différents niveaux d'expertise. Cette approche est conforme aux pratiques de la science ouverte et pourrait améliorer la qualité de la documentation.[59] En liaison avec la mesure 21 relative à la documentation, le site[60] pourrait encourager l'utilisation de ces formats afin de promouvoir l'établissement de rapports complets et transparents tout au long du cycle de vie de l'IA.

4.9 Automatiser les décisions de développement et de déploiement

La mesure 14 stipule que "[l]es ignares s'engagent à mettre en place un processus permettant de décider s'il convient ou non de poursuivre le développement et le déploiement d'un modèle d'IA à usage général présentant un risque systémique".[61]

L'automatisation de certains aspects de ce processus décisionnel pourrait améliorer l'objectivité et la cohérence. Par exemple, des critères ou des seuils prédéfinis, tels qu'ils sont énumérés dans d'autres parties du code, pourraient déclencher des actions automatisées, telles que l'arrêt du développement si certains indicateurs de risque dépassent les niveaux acceptables. Cette approche permettrait d'éviter les biais ou les pressions qui pourraient conduire les développeurs à négliger des risques potentiels, à l'instar de ce que l'on appelle le "p-hacking" dans la recherche scientifique.[62] Le piratage fait référence à la manipulation des méthodes d'analyse des données ou à la communication sélective des résultats pour produire artificiellement des résultats statistiquement significatifs, souvent en exploitant les degrés de liberté des chercheurs, ce qui peut conduire à des faux positifs et porter atteinte à l'intégrité de la recherche scientifique. Cependant, l'automatisation complète n'est pas toujours possible en raison de la complexité des considérations éthiques et de la nécessité d'un jugement humain. Une approche hybride, dans laquelle l'automatisation soutient mais ne remplace pas la prise de décision humaine, peut s'avérer plus efficace. En liaison avec la mesure 18 relative à la notification des incidents graves,[63] , un système intégré pourrait automatiquement signaler les incidents et garantir leur notification en temps utile à l'Office de l'IA de l'Union européenne et à d'autres autorités compétentes.

4.10 Définition des incidents graves

La mesure 18 pose la question : "Qu'est-ce qu'un incident grave ?" et définit les obligations en matière de suivi, de documentation et de notification de ces incidents.[64]

Une méthode judicieuse pour définir les incidents graves pourrait consister à adopter un protocole de pré-enregistrement, semblable à celui utilisé dans les essais cliniques, dans lequel les résultats et les risques attendus sont documentés à l'avance, créant ainsi une division entre la "prédiction" et la "postdiction".[65] Les écarts par

---

[58] Thomas Kluyver et al, 'Jupyter Notebooks ? A Publishing Format for Reproducible Computational Workflows', in *Positioning and Power in Academic Publishing : Players, Agents and Agendas*, ed. Fernando Loizides et Birgit Scmidt (IOS Press, 2016), 87-90, https://eprints.soton.ac.uk/403913/.

[59] Jeffrey M. Perkel, " Why Jupyter Is Data Scientists' Computational Notebook of Choice ", *Nature* 563, no 7729 (novembre 2018) : 145–46, https://doi.org/10.1038/d41586-018-07196-1.

[60] Code de conduite, 1er projet, p. 34.

[61] Code de conduite, 1er projet, p. 29.

[62] Megan L. Head et al, "The Extent and Consequences of P-Hacking in Science", *PLOS Biology* 13, no. 3 (13 mars 2015) : e1002106, https://doi.org/10.1371/journal.pbio.1002106.

[63] Code de conduite, 1er projet, p. 32.

[64] Code de conduite, 1er projet, p. 32.

[65] Brian A. Nosek et al, "The Preregistration Revolution", *Proceedings of the National Academy of Sciences* 115, no 11 (13 mars 2018) : 2600-2606, https://doi.org/10.1073/pnas.1708274114.

rapport aux résultats escomptés pourraient alors être objectivement identifiés comme des événements indésirables graves potentiels. Par ailleurs, l'intégration d'un jugement d'expert a posteriori, par exemple par l'intermédiaire d'un comité d'évaluation éthique, pourrait aider à déterminer la gravité des incidents.[66]

4.11 Sanctions en cas d'absence de notification

La mesure 20.1 permet à la Commission de désigner un modèle comme présentant un risque systémique si elle a connaissance d'un tel modèle qui n'a pas été notifié.[67] Toutefois, rien n'indique qu'il existe une disposition prévoyant des sanctions concrètes en cas de défaut de notification.

L'introduction de sanctions pour défaut de notification des modèles GPAI pertinents à l'Office de l'IA pourrait accroître la conformité, en particulier parmi les grands fournisseurs d'IA susceptibles d'avoir un impact significatif sur le marché. On pourrait s'inspirer du droit européen de la concurrence, où le fait de ne pas notifier les fusions peut entraîner des amendes importantes, qui ont un effet dissuasif.[68] Toutefois, l'imposition de sanctions doit être soigneusement étudiée afin d'éviter d'affecter de manière disproportionnée les PME, les développeurs de logiciels libres ou les chercheurs universitaires qui peuvent manquer de ressources ou d'expertise juridique. Une exemption pour ces groupes est justifiée afin de reconnaître leurs capacités différentes et d'encourager l'innovation. Les sanctions devraient se concentrer sur les entreprises qui ont la capacité et la responsabilité de se conformer mais qui choisissent de ne pas le faire. Des campagnes de sensibilisation devraient permettre de s'assurer que tous les fournisseurs comprennent leurs obligations.

4.12 Critères de classification dynamique au-delà des seuils de calcul

Le code indique que "[l]'Office AI est habilité à mettre à jour les critères de classification permettant de déterminer si un modèle à usage général est présumé avoir des capacités à fort impact [...]. Comment pourrait-on rédiger cette disposition de manière à ce qu'il soit clair à quel moment les fournisseurs doivent notifier à l'Office AI qu'un modèle répond aux nouveaux critères de classification ?[69]

S'appuyer uniquement sur des seuils de calcul pour la classification (par exemple, des modèles formés avec plus de 10^25 opérations en virgule flottante) est de plus en plus inadéquat, comme l'ont récemment fait remarquer des chercheurs.[70] Les progrès de l'efficacité algorithmique et des méthodes d'apprentissage décentralisées signifient que des modèles puissants peuvent être développés sans dépasser des seuils de calcul élevés. D'autres critères pourraient inclure des repères de performance dans des tests standardisés ou des évaluations de compétences. Par exemple, en se référant à des tableaux de classement en constante évolution tels que le "Chatbot Arena", ou à des critères de référence de la communauté AI Alignment,[71] peut fournir des évaluations plus dynamiques et contextuelles. Il est essentiel que l'Office de l'IA communique clairement sur les mises à jour des critères de classification.

---

66 Voir aussi : Luciano Floridi et al, 'AI4People-An Ethical Framework for a Good AI Society : Opportunities, Risks, Principles, and Recommendations', *Minds and Machines* 28, no. 4 (décembre 2018) : 689–707, https://doi.org/10.1007/s11023-018-9482-5.

67 Code de conduite, 1er projet, p. 33.

68 Ariel Ezrachi, *EU Competition Law : An Analytical Guide to the Leading Cases*, 5e édition (Oxford : Hart, 2016).

69 Code de conduite, 1er projet, p. 34.

70 Sara Hooker, "On the Limitations of Compute Thresholds as a Governance Strategy" (arXiv, 2024), https://doi.org/10.48550/ARXIV.2407.05694.

71 Dan Hendrycks et al, "Aligning AI With Shared Human Values" (arXiv, 2020), https://doi.org/10.48550/ARXIV.2008.02275.

5 Conclusion

Le code de bonnes pratiques en matière d'IA à usage général aidera l'UE à orienter les développeurs de modèles d'IA vers la promotion d'un écosystème d'IA sûr, digne de confiance et innovant en Europe. Cette analyse, basée sur le premier projet accessible au public, souligne les points forts du code, mais aussi les domaines dans lesquels des améliorations pourraient renforcer son efficacité. Tout d'abord, en prévoyant une certaine souplesse pour s'adapter aux avancées technologiques, le code reconnaît l'évolution rapide de l'IA et la nécessité d'une réglementation dynamique. Le principe de proportionnalité garantit que les obligations sont équilibrées par rapport aux risques et aux capacités des fournisseurs, en soutenant notamment les PME et les jeunes entreprises. En outre, la reconnaissance explicite des contributions des fournisseurs de logiciels libres renforce le développement de l'IA à long terme et favorise l'inclusion.

Deuxièmement, les dispositions spécifiques relatives au groupe de travail 2 soulignent l'importance d'une approche globale du risque systémique. Il serait essentiel d'élargir la taxonomie pour y inclure les effets en cascade et les retombées afin d'appréhender des scénarios de risque complexes, en particulier dans les systèmes interconnectés au cours d'un scénario de "polycrise". L'intégration de méthodologies interdisciplinaires enrichit l'analyse des risques en fournissant une compréhension multidimensionnelle et plus holistique des menaces potentielles. L'établissement de niveaux de gravité clairs permet de hiérarchiser les efforts d'atténuation, tandis que l'assurance d'une grande rigueur scientifique dans les évaluations maintient la confiance et la responsabilité. L'évaluation continue des risques tout au long du cycle de vie de l'IA, en particulier lors du déploiement, permet de prendre en compte les risques émergents liés à l'utilisation dans le monde réel et aux interactions entre l'homme et l'IA.

Enfin, ce document de retour d'information présente quelques recommandations pratiques pour affiner le code. La mise en place d'un processus transparent de révision et de mise à jour, avec des voies accessibles pour la contribution de la société civile, pourrait garantir que le code reste à jour et reflète des perspectives diverses. Un meilleur équilibre des exigences de transparence pour les fournisseurs en aval permet d'éviter les charges excessives pour les PME et encourage une participation plus large à l'innovation dans le domaine de l'intelligence artificielle. L'intégration de formats innovants pour les rapports de sûreté et de sécurité, tels que les carnets Jupyter, améliorerait la clarté et faciliterait une meilleure compréhension entre les parties prenantes. L'automatisation de certains aspects du processus de prise de décision en matière de développement et de déploiement pourrait améliorer l'objectivité et éviter les préjugés. Des définitions claires des incidents graves et des sanctions proportionnées en cas de non-conformité sont essentielles pour un cadre réglementaire solide.

En conclusion, le premier projet constitue une base solide pour le code de pratique polyvalent en matière d'IA. Le code sera finalisé d'ici avril 2025, soit neuf mois après l'entrée en vigueur de la loi sur l'IA, le 1er août 2024. En abordant les domaines d'amélioration mis en évidence, le code pourrait équilibrer plus efficacement la promotion de l'innovation et l'impératif de sécurité. Cela favorisera un paysage européen de l'IA qui non seulement stimule le progrès technologique, mais protège également les valeurs sociétales et atténue les risques systémiques. Le Centre for European Policy (cep) continuera à apporter son expertise, en veillant à ce que le code serve au mieux les intérêts de la société.

Centrum für Europäische Politik
FREIBURG | BERLIN

Auteur :

Anselm Küsters, LL.M., chef de la division Numérisation et nouvelles technologies
kuesters@cep.eu

Traductrice:

Emma Drouet, chargée de communication & de relations presses au cep
drouet@cep.eu

Centrum für Europäische Politik FREIBURG | BERLIN
Kaiser-Joseph-Straße 266 | D-79098 Freiburg
Schiffbauerdamm 40 Räume 4205/06 | D-10117 Berlin
Tél. + 49 761 38693-0

Le **Centrum für Europäische Politik** FREIBURG | BERLIN, le **Centre de Politique Européenne** PARIS, et le **Centro Politiche Europee** ROMA forment le **réseau des Centres de Politique Européenne** FREIBURG | BERLIN | PARIS | ROMA.

Exempt d'intérêts particuliers et neutre sur le plan politique, le réseau des centres de politique européenne fournit une analyse et une évaluation de la politique de l'Union européenne, dans le but de soutenir l'intégration européenne et de défendre les principes d'un système économique de libre-échange.