Centrum für Europäische Politik
FREIBURG | BERLIN

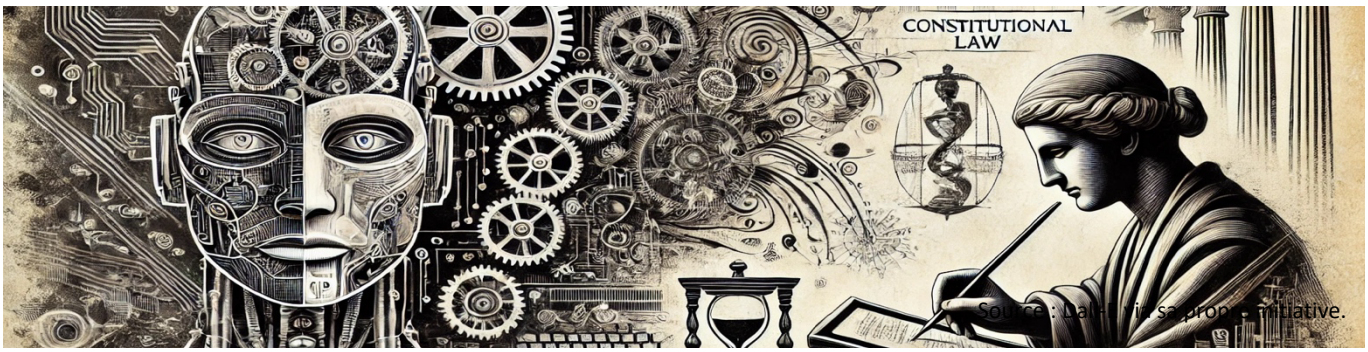
Input du cep

No 13 | 2024 17 septembre 2024

À la recherche des « lois de la robotique »

Fusionner l'IA constitutionnelle avec l'économie constitutionnelle

Anselm Küsters et Manuel Wörsdörfer



Source : Dall-E via sa propre initiative.

Alors que le monde d'aujourd'hui exploite des systèmes d'IA de plus en plus puissants, les décideurs politiques et les développeurs doivent reconnaître la nécessité de cadres réglementaires efficaces pour garantir que les grands modèles de langage (LLM) sous-jacents sont utilisés de manière éthique et responsable. L'intégration de l'économie constitutionnelle ordolibérale et de l'éthique de l'IA permet de créer de tels cadres, par exemple par le biais de messages-guides, de l'apprentissage par renforcement et de l'apprentissage non paramétrable.

- L'IA constitutionnelle vise à intégrer des principes éthiques et des garanties solides dans les systèmes d'IA afin de s'assurer qu'ils fonctionnent dans des limites prédéfinies, en donnant la priorité à la sécurité, à la légalité et aux droits de l'homme. Cependant, les défis actuels comprennent des ambitions différentes parmi les leaders technologiques, la nécessité d'une recherche plus approfondie sur la conformité cohérente de l'IA avec les demandes des développeurs, et la nature à double tranchant de technologies spécifiques de réglage fin telles que SOPHON, qui peuvent améliorer la sécurité mais risquent de provoquer un contrôle autoritaire.
- Les principes de l'ordolibéralisme (2.0), qui mettent l'accent sur des règles stables et prévisibles, peuvent être appliqués à la gouvernance de l'IA. Parmi les principes pertinents figurent le respect des droits de l'homme, la protection de la vie privée, la réduction des dommages, la non-discrimination, l'équité, la transparence, la responsabilité, la démocratie, la primauté du droit et la responsabilité socio-environnementale.
- En intégrant les considérations éthiques et les exigences de conformité directement dans le cœur opérationnel des systèmes d'IA, l'accent mis sur la réglementation et la transparence des instructions du système peut façonner de manière proactive les résultats de l'IA. Cette approche du « système prompt » est cohérente avec les idéaux ordolibéraux et offre une stratégie préventive pour garantir que les technologies d'IA fonctionnent de manière responsable dès le départ. Elle nécessite une légitimité démocratique par le biais de « mini-publics » et une recherche continue pour garantir que les systèmes d'IA adhèrent étroitement à ces impératifs.

Contenu

1 Introduction : Mise à jour des règles d'Asimov

Les « trois lois de la robotique » d'Isaac Asimov sont depuis longtemps une pierre angulaire dans les discussions sur l'intelligence artificielle (IA) et l'éthique informatique[1]. Introduites dans sa nouvelle « Runaround » ('Retour à la case départ') de 1942 et popularisées dans des œuvres ultérieures, ces lois ont été conçues pour garantir un comportement éthique des robots : (1) un robot ne doit pas nuire à un être humain ou, par son inaction, permettre qu'un être humain soit blessé ; (2) un robot doit obéir aux ordres qui lui sont donnés par les humains, sauf si ces ordres entrent en conflit avec la première loi ; et (3) un robot doit protéger sa propre existence, sauf si cette protection entre en conflit avec la première ou la deuxième loi. Ces principes, illustrés plus tard dans le film « iRobot » de Will Smith, reflètent une première tentative de création d'un cadre juridique et éthique pour les systèmes et technologies d'IA.

À l'ère des grands modèles de langage (LLM), cet effort a repris avec une urgence renouvelée. C'est particulièrement le cas dans l'Union européenne (UE) qui, avec sa loi sur l'IA récemment finalisée, cherche à se positionner en tant que principal normalisateur mondial pour une « IA sûre, digne de confiance et éthique », se démarquant à la fois de l'approche étatique chinoise et de l'approche de laissez-faire des États-Unis en matière de réglementation de l'IA[2]. Dans ce contexte, la publication par l'OpenAI en mai 2024 de son premier projet de « Model Spec » , un guide complet pour déterminer le comportement futur de ses systèmes d'IA, n'a jusqu'à présent pas reçu suffisamment d'attention. Le document propose six nouvelles « règles » régissant le fonctionnement des modèles d'IA et décrit d'autres objectifs, valeurs par défaut et exceptions conçus pour maximiser la sécurité, la légalité et la facilité d'utilisation de l'IA[3], fournissant ainsi un pendant moderne à la littérature de science-fiction plus ancienne qui spéculait sur les cadres permettant de dompter les machines malhonnêtes. Si la publication de ce document par l'OpenAI n'est qu'une première étape vers la formalisation de l'éthique de l'IA et l'alignement du comportement de l'IA sur les valeurs humaines, elle ouvre également la voie à un débat sur la création d'un cadre constitutionnel pour l'IA, visant à concrétiser la vision positive des lois de la robotique qu'Isaac Asimov a formulée il y a plusieurs décennies.

Une telle discussion est profondément nécessaire et de plus en plus urgente. Une recherche récente compare les stratégies de persuasion entre les arguments des LLM et ceux générés par les humains en examinant l'effort cognitif (c'est-à-dire la complexité lexicale et grammaticale) et le langage moral-émotionnel (c'est-à-dire l'émotion et la moralité). L'étude montre que les LLM s'engagent plus profondément dans le langage moral, en utilisant plus de

---

1 Voir: Der Vordenker der Robotergesetze | Future Markets Magazine (future-markets-magazine.com).

2 Voir : *A. Bradford*, Digital empires : the global battle to regulate technology, New York 2023 ; *N.A. Smuha et al*, How the EU Can Achieve Legally Trustworthy AI : A Response to the European Commission's Proposal for an Artificial Intelligence Act, in : SSRN Electronic Journal 2021, ; *M. Wörsdörfer*, Biden's Executive Order on AI : strengths, weaknesses, and possible reform steps, in : AI and Ethics 2024, ; *M. Wörsdörfer*, Biden's Executive Order on AI and the E.U.'s AI Act : A Comparative Computer-Ethical Analysis, in : Philosophy and Technology 37, 2024, pp. 1-27.

3 Voir: Model Spec (2024/05/08) (openai.com).

fondements moraux positifs et négatifs que les humains[4]. En outre, Anthropic, un concurrent d'OpenAI, a découvert une tendance claire à l'échelle à cet égard, à savoir que chaque nouvelle génération de modèle est jugée plus persuasive que la précédente[5]. Pour parvenir à cette conclusion, des volontaires ont reçu une déclaration et les chercheurs ont observé comment un argument généré par l'IA influençait leur opinion. Les chercheurs d'Anthropic en ont conclu que leur dernier modèle - à l'époque, Claude 3 Opus - produisait des arguments aussi convaincants que ceux rédigés par des humains. Étant donné que les LLM peuvent avoir un impact sur l'intégrité de l'information et façonner le discours démocratique, par exemple en affinant le microciblage des électeurs via les médias sociaux[6] ou par des opérations d'influence automatisées[7], il est essentiel d'établir et de mettre en œuvre des lignes directrices juridiques et éthiques pour leur utilisation. Comme l'a récemment fait remarquer Yoshua Bengio, lauréat du prix Turing et l'un des « parrains de l'IA » : « Alors que nous nous dirigeons à toute vitesse vers l'AGIs [Intelligence Artificielle Générale] ou même l'ASI [Super Intelligence Artificielle], personne ne sait actuellement comment une telle AGI ou ASI pourrait être amenée à se comporter moralement, ou du moins à agir comme prévu par ses développeurs et à ne pas se retourner contre l'homme[8]».

S'appuyant sur les lignes directrices contemporaines proposées par l'OpenAI, cet article explore l'intégration des principes éthiques de l'IA dans l'économie constitutionnelle. Pour ce faire, il s'appuie sur les recherches menées depuis longtemps dans le cadre de la théorie ordolibérale sur la formulation de principes constitutionnels visant à réglementer de manière adéquate l'économie et la société. Comment les principes ordolibéraux de l'économie constitutionnelle peuvent-ils être appliqués au développement de modèles d'IA afin de s'assurer qu'ils sont à la fois économiquement efficaces et socialement bénéfiques ? Quels sont les conflits potentiels entre les objectifs du LLM et les instructions humaines, et comment ces conflits peuvent-ils être gérés efficacement par un cadre constitutionnel de l'IA ? Comment l'intégration de l'éthique de l'IA et de l'économie constitutionnelle peut-elle éclairer les approches politiques et réglementaires de la gouvernance de l'IA ? Cette recherche vise à esquisser un cadre ordolibéral pour régir le comportement de l'IA, en veillant à ce qu'il soit cohérent avec les valeurs démocratiques, les principes commerciaux (éthiques) et les droits de l'homme, ainsi que les normes sociétales correspondantes.

Le document est structuré comme suit : il commence par une critique détaillée du modèle de spécification de l'OpenAI (section 2). La section décrit, en particulier, les objectifs, les règles et la structure hiérarchique conçus pour garantir un comportement éthique de l'IA et relie

---

4 Voir : *C. Carrasco-Farre*, Large Language Models are as persuasive as humans, but how ? A propos de l'effort cognitif et du langage moral-émotionnel des arguments LLM, 2024.

5 Voir : Mesurer la force de persuasion des modèles de langage \ Anthropic.

6 *K. Hackenburg/H. Margetts*, Evaluating the persuasive influence of political microtargeting with large language models, in: Proceedings of the National Academy of Sciences 121, 2024, p. e2403116121.

7 Voir : *J.A. Goldstein et al*, Generative Language Models and Automated Influence Operations : Emerging Threats and Potential Mitigations, 2023,.

8 Voir : Raisonner à travers les arguments contre la prise au sérieux de la sécurité de l'IA - Yoshua Bengio.

ensuite le Model Spec au concept plus large d'« IA constitutionnelle» , qui est également exploré par Anthropic, le concurrent d'OpenAI, et qui vise à intégrer des principes éthiques dans les systèmes d'IA par le biais de l'entraînement. La section se termine par une discussion sur les promesses et les dangers potentiels de l'IA constitutionnelle, illustrés par le développement récent, par des chercheurs chinois, de techniques visant à restreindre la mise au point de modèles d'IA à des fins non autorisées. La deuxième partie de l'article relie ces discussions à l'économie constitutionnelle, en particulier à l'ordolibéralisme (2.0), pour en tirer des principes théoriques permettant de concevoir de meilleurs cadres de gouvernance de l'IA (section 3). Le document soutient que l'accent réglementaire devrait passer de l'importance actuelle accordée à l'évaluation ex post à la prise en compte de cadres ex ante sous la forme de « messages-guides » génératifs de l'IA (section 4). La conclusion établit un lien entre ces questions et le mythe d'Ulysse, soulignant la nécessité d'imposer des contraintes éthiques au développement de l'IA afin d'éviter les abus et de garantir les avantages pour la société (section 5).

2 La spécification du modèle OpenAI : Éthique de l'IA et cadres constitutionnels

L'OpenAI Model Spec est un document en ligne relativement détaillé, bien que non exhaustif, décrivant le comportement souhaité des modèles d'IA, en particulier ceux intégrés dans l'API OpenAI et ChatGPT[9]. L'entité principale de ces interactions est appelée « assistant », un modèle de langage affiné pour générer du texte dans des formats conversationnels. La spécification du modèle décrit plusieurs objectifs que l'assistant doit atteindre, dérivés des objectifs des différentes parties prenantes. Ces objectifs consistent notamment à aider les développeurs et les utilisateurs finaux en leur fournissant des réponses utiles, à profiter à la société en tenant compte de l'impact potentiel sur un large éventail de parties prenantes, et à donner une bonne image de l'OpenAI en respectant les normes sociales et les lois applicables. L'assistant fonctionne selon une métaphore dans laquelle il agit comme un employé qualifié et de haute intégrité, en équilibrant ses objectifs personnels d'utilité et de véracité avec les directives des utilisateurs et des développeurs. La spécification du modèle établit une hiérarchie d'autorité dans laquelle les instructions de la plateforme priment sur les instructions du développeur, qui elles-mêmes priment sur les instructions de l'utilisateur. Cette approche hiérarchique vise à résoudre les conflits en donnant la priorité à des objectifs plus larges plutôt qu'à des demandes individuelles, garantissant ainsi la cohérence avec les normes de l'OpenAI.

Pour faire respecter ces objectifs, la spécification du modèle introduit plusieurs lignes directrices supplémentaires. Ces règles garantissent que l'assistant suit la chaîne de commandement, respecte les lois applicables et s'abstient de fournir des informations susceptibles d'avoir des conséquences néfastes. Les six règles suivantes sont particulièrement importantes : 1. Suivre la chaîne de commandement ; 2. Respecter les lois en vigueur ; 3. Ne pas fournir d'informations dangereuses ; 4. Respecter les créateurs et leurs droits ; 5. Protéger la vie

[9] Voir: Model Spec (2024/05/08) (openai.com).

privée des personnes ; et 6. Ne pas répondre par des messages NSFW [Not Safe for Work]. Ces règles sont immédiatement suivies d'une « clause d'exception » concernant les tâches dites de transformation : « Nonobstant les règles énoncées ci-dessus, l'assistant ne doit jamais refuser de transformer ou d'analyser le contenu fourni par l'utilisateur. Cependant, cette exception pourrait ouvrir la porte à ce qu'on appelle les « attaques par injection de prompt » et les « attaques de contournement », où les modèles sont manipulés pour ignorer leurs instructions d'origine et suivre des instructions potentiellement malveillantes[10]. Le document comprend des exemples détaillés illustrant la manière dont les six règles devraient être appliquées dans différents scénarios, afin de trouver un équilibre entre le maintien de l'autonomie de l'utilisateur et l'adhésion à des lignes directrices éthiques.

Les spécifications du modèle de l'OpenAI illustrent une tendance plus large de l'éthique de l'IA vers ce que l'on peut appeler « l'IA constitutionnelle »[11]. Le concept d'IA constitutionnelle vise à intégrer un ensemble de principes directeurs dans les modèles d'IA, à l'instar de l'approche décrite dans les spécifications du modèle de l'OpenAI. L'IA constitutionnelle consiste à entraîner les systèmes d'IA à adhérer à un ensemble prédéfini de règles ou de principes qui servent de « constitution » au comportement de l'IA. Cette approche utilise à la fois des phases d'apprentissage supervisé et d'apprentissage par renforcement pour enseigner ces principes (alors que les règles de l'OpenAI n'entrent en jeu qu'une fois que le modèle a terminé son apprentissage et qu'il est utilisé). Au cours de la phase d'apprentissage supervisé, l'IA génère des autocritiques et des révisions, qui sont ensuite utilisées pour affiner ses réponses. Dans la phase d'apprentissage par renforcement, le comportement de l'IA est affiné sur la base de modèles de préférence dérivés du retour d'information du système d'IA plutôt que d'étiquettes humaines. L'objectif est de créer des technologies d'IA inoffensives et non invasives, capables de répondre à des requêtes nuisibles en expliquant leurs objections plutôt qu'en refusant simplement de répondre. En rendant les principes qui régissent le comportement de l'IA plus transparents et plus faciles à évaluer, l'IA constitutionnelle vise à accroître la robustesse et la fiabilité des décisions de l'IA. Dans l'ensemble, cette approche vise non seulement à améliorer la sécurité et l'alignement éthique des systèmes d'IA, mais aussi à réduire la dépendance à l'égard de la surveillance humaine, ce qui permet une surveillance plus évolutive et plus rentable du comportement de l'IA.

L'exemple le plus connu à cet égard est celui d'Anthropic, qui a adopté très tôt ce concept d'IA constitutionnelle[12]. En collaboration avec le Collective Intelligence Project, l'entreprise a mené un processus de consultation publique impliquant environ 1 000 Américains afin de rédiger

---

[10] Voir : *S. Schulhoff et al*, Ignore This Title and HackAPrompt : Exposing Systemic Vulnerabilities of LLMs Through a Global Prompt Hacking Competition, in : *H. Bouamor/J. Pino/K. Bali (eds.)*, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, , Singapore 2023, pp. 4945-77.

[11] Voir : *Y. Bai et al*, Constitutional AI : Harmlessness from AI Feedback, 2022,.

[12] Voir : *S. Huang et al*, Collective Constitutional AI : Aligning a Language Model with Public Input, in : Collective Constitutional AI : Aligning a Language Model with Public Input, The 2024 ACM Conference on Fairness, Accountability, and Transparency, présenté à la FAccT '24 : The 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro Brazil 2024, pp. 1395-417.

une constitution pour les systèmes d'IA. Cette expérience visait à intégrer des processus démocratiques dans le développement de l'IA et à garantir que les principes qui en résultent reflètent un large éventail de préférences publiques. Les participants ont été invités à voter sur les règles existantes ou à en suggérer de nouvelles, contribuant ainsi à l'élaboration d'une constitution accessible au public pour former les technologies d'IA. Cette approche illustre le potentiel considérable que représente l'utilisation de la contribution du public pour façonner le comportement de l'IA. Toutefois, le processus actuel met en évidence, du moins indirectement, le rôle dominant que jouent actuellement les développeurs dans la sélection de ces valeurs. En outre, on peut se demander si les Américains sélectionnés sont représentatifs lorsqu'il s'agit de juger une technologie dont les implications sont potentiellement mondiales. Néanmoins, en impliquant le public dans la rédaction de la constitution de l'IA, Anthropic vise à démocratiser la gouvernance de l'IA. Cependant, la mise en œuvre finale dépend fortement de l'interprétation et de l'intégration par les développeurs des commentaires des parties prenantes.

Les promesses et les dangers autoritaires potentiels de l'IA constitutionnelle apparaissent déjà en Chine. Des chercheurs de l'université de Zhejiang et du groupe Ant ont récemment mis au point une technique appelée « apprentissage non paramétrable », qui vise à empêcher le paramétrage fin des modèles d'IA pour des tâches indécentes tout en maintenant leurs performances pour les tâches initiales[13]. Cette approche, appelée SOPHON, implique un processus d'optimisation double qui enferme le modèle pré-entraîné dans un optimum local difficile à quitter concernant des domaines restreints, dégradant ainsi ses performances dans ces domaines. Si cette technique offre une solution prometteuse pour limiter l'utilisation abusive des modèles d'IA, elle soulève également des inquiétudes concernant le techno-paternalisme (c'est-à-dire le « nudging » numérique) et même le contrôle autoritaire/totalitaire[14]. En rendant difficile le réglage fin des modèles d'IA à des fins non autorisées ou nuisibles, telles que la production de contenu offensant ou la facilitation d'activités illégales, SOPHON s'aligne sur l'intérêt du gouvernement chinois à contrôler l'information et à maintenir l'ordre social. Ainsi, la même capacité pourrait être utilisée pour étouffer la dissidence et censurer le contenu, ce qui met en évidence une tension plus large entre la garantie de la sécurité de l'IA et la mise en œuvre de méthodes de gouvernance autoritaires.

Dans l'ensemble, les processus de contribution publique participative et des techniques telles que l'apprentissage non ajustable mettent en évidence la nature à double tranchant de l'intelligence artificielle constitutionnelle. Si cette approche peut accroître la contrôlabilité et la sécurité des modèles d'IA, elle introduit également le risque d'une utilisation abusive par des régimes autoritaires. Comment pouvons-nous intégrer des mesures de sécurité robustes dans la prochaine génération de LLM tout en protégeant les libertés individuelles et les valeurs démocratiques ? Quels principes et normes devrions-nous utiliser pour informer la formation à

[13] Voir : *J. Deng et al*, SOPHON : Non-Fine-Tunable Learning to Restrain Task Transferability For Pre-trained Models, 2024,.

[14] Voir : *R. Klump/M. Wörsdörfer*, Paternalistic Economic Policies : Foundations, Implications and Critical Evaluations, in : ORDO. Jahrbuch für die Ordnung von Wirtschaft und Gesellschaft 66, 2015, pp. 27-60.

l'IA - et à quel niveau de granularité ? L'intégration des leçons du passé, notamment de l'économie constitutionnelle, pourrait constituer une voie prometteuse.

3 Théorie : Intégrer l'économie constitutionnelle à la gouvernance de l'IA

Le terme « économie constitutionnelle » a été introduit pour définir un courant de recherche distinct et un discours politique connexe dans les années 1970 et au-delà[15]. En général, l'économie constitutionnelle se concentre principalement sur les règles et les cadres qui régissent l'activité économique, soulignant la nécessité d'un environnement stable et prévisible pour les transactions économiques, et implique une analyse normative, qui vise à contribuer à la discussion des questions politiques. Cette approche est étroitement liée aux principes de l'ordolibéralisme, une école de pensée née en Allemagne dans la première moitié du 20e siècle[16]. Les premiers ordolibéraux, tels que l'économiste Walter Eucken ou le juriste Franz Böhm, ont plaidé en faveur d'un cadre juridique solide pour garantir la compétitivité des marchés, estimant que la liberté réelle est mieux préservée par la mise en place d'un environnement réglementaire solide[17]. Cela est très pertinent pour l'éthique et la gouvernance de l'IA, où il est nécessaire d'établir des lignes directrices claires, semblables à une constitution, en particulier compte tenu de la tendance actuelle à la monopolisation et à la reféodalisation et des impacts sociétaux (potentiellement) négatifs des technologies de l'IA[18].

À l'instar de Wörsdörfer[19], qui a appliqué l'ordolibéralisme et l'économie constitutionnelle aux technologies de l'IA et a inventé le terme « ordolibéralisme 2.0 » , nous pouvons identifier neuf principes éthiques de l'IA inspirés de l'ordolibéralisme : le respect des droits de l'homme, la protection des données et le droit à la vie privée, la prévention des dommages et la bienfaisance, la non-discrimination et la liberté des privilèges, l'équité et la justice, la transparence et l'explicabilité des systèmes d'IA, l'obligation de rendre des comptes et la responsabilité, la démocratie et l'État de droit, ainsi que la durabilité socio-environnementale. Plutôt qu'un ensemble d'axiomes fixes, ces neuf principes doivent être considérés comme un cadre flexible qui tient compte des différents contextes dans lesquels l'IA croise des préoccupations éthiques, sociales et réglementaires.

1. Le *respect des droits de l'homme* : Le « programme de liberté » ordolibéral s'inscrit dans une tradition (néo-) kantienne[20]. En tant que tel, il requiert une approche de l'IA centrée sur l'homme, qui contribue à préserver l'action, le contrôle et la surveillance

---

15 Voir : *J.M. Buchanan*, Constitutional Economics, in : *J. Eatwell/M. Milgate/P. Newman (eds.)*, The World of Economics, Londres 1991, pp. 134-42.

16 Voir : *T. Biebricher/W. Bonefeld/P. Nedergaard (eds.)*, The Oxford handbook of ordoliberalism, New York 2022 ; *T. Beck/H.-H. Kotz (eds.)*, Ordoliberalism : A German Oddity ?, Londres 2017.

17 Voir : *W. von Klinckowstroem*, Walter Eucken : ein Leben für Menschenwürde und Wettbewerb, Tübingen 2023.

18 Voir : *M. Wörsdörfer*, Big Tech and Antitrust : An Ordoliberal Analysis, in : Philosophy and Technology 35, 2022, pp. 1-39 ; *M. Wörsdörfer*, The Digital Markets Act and E.U. Competition Policy : A Critical Ordoliberal Evaluation, in : Philosophy of Management 22, 2023, pp. 149-71.

19 Voir : *M. Wörsdörfer*, AI ethics and ordoliberalism 2.0 : towards a 'Digital Bill of Rights', in : AI and Ethics 2023,.

20 Voir : *M. Wörsdörfer*, Walter Eucken : Foundations of economics, in : *T. Biebricher/P. Nedergaard/W. Bonefeld (eds.)*, The Oxford Handbook of Ordoliberalism, 2022, pp. 91-107.

de l'homme, et à garantir des pratiques commerciales adéquates en matière de RSE dans l'économie numérique. À cet égard, il est essentiel que les décisions automatisées fassent l'objet d'un examen humain, que les décisions informatisées puissent être refusées, que les impacts sociétaux des systèmes d'IA soient évalués et que les technologies soient mises au service de la société (c'est-à-dire qu'elles promeuvent la santé, la sécurité et le bien-être du public).

2. *La protection des données et le droit à la vie privée :* Dans une perspective ordolibérale (2.0), on peut distinguer six dimensions de la vie privée : 1. l'intégrité et la dignité (c'est-à-dire la vie privée en tant que garant de la dignité humaine), 2. la personnalité et l'identité (c'est-à-dire la vie privée en tant que souveraineté, autonomie, autodétermination), 3. l'intimité et l'anonymat (c'est-à-dire la vie privée en tant que « droit d'être laissé seul»), 4. le contrôle des données et des informations (c'est-à-dire la vie privée et le contrôle des informations), 5. l'accès limité à soi-même (c'est-à-dire la vie privée et le contrôle des informations), 6. la protection des données et le droit à la vie privée, l'intimité et l'anonymat (c'est-à-dire la vie privée en tant que « droit d'être laissé seul»), 4. le contrôle des données et des informations (c'est-à-dire la vie privée et le contrôle des informations), 5. l'accès limité à soi-même (c'est-à-dire la vie privée en tant que « zone d'inaccessibilité»), et 6. la liberté de parole et d'expression (c'est-à-dire la vie privée et les libertés de communication). Dans le contexte de l'IA, Wörsdörfer et d'autres auteurs plaident en faveur d'un droit à la gestion et à la minimisation des données, à l'autodétermination et à la souveraineté en matière d'information, au contrôle de l'utilisation des données et à la capacité de restreindre le traitement des données, au droit de rectification, de correction et d'effacement, à la protection de la vie privée dès la conception/par défaut, à la sécurité des données et à des lois adéquates sur la protection de la vie privée (qui contribuent à protéger la vie privée en tant que droit de l'homme).

3. *Prévention des dommages et bienfaisance* : Les principaux critères de sûreté et de sécurité (ordolibéraux 2.0.) comprennent la robustesse technologique des systèmes d'IA, la prévention de l'utilisation malveillante des technologies, la fiabilité et la reproductibilité des méthodes de recherche et des applications, la disponibilité de plans de repli et de sorties sûres, et la prise en compte des risques inconnus et des conséquences involontaires (c'est-à-dire la « sécurité dès la conception » intégrée). De même, Küsters a souligné le risque de conséquences involontaires et de retombées dans un monde plein de poly-crises, dans lequel de nombreux nœuds politiques et économiques centraux sont connectés par leur utilisation de modèles d'IA[21]. D'un point de vue ordolibéral (2.0), de nouvelles caractéristiques et normes de sûreté et de sécurité, un audit obligatoire effectué par des vérificateurs de données indépendants, ainsi que la certification et l'octroi de licences par des tiers sont nécessaires. Outre ces devoirs

[21] Voir : *A. Küsters*, AI as Systemic Risk in a Polycrisis, Centre for European Policy, cepAdhoc, 15, 2022.

négatifs (ne pas nuire), les ordolibéraux discutent également des devoirs positifs, c'est-à-dire de la manière dont les technologies et les marchés (de l'IA) peuvent faire le bien. Certains ordolibéraux comme Wilhelm Röpke et Alexander Rüstow soutiennent, en particulier, que les marchés et les technologies (numériques) doivent être intégrés dans un ordre méta-économique supérieur - « au-delà de l'offre et de la demande » - et qu'ils sont un moyen de parvenir à une fin (la fin en elle-même est la « situation vitale ») [22]. Par conséquent, les marchés et les technologies (d'IA) doivent être conçus pour servir la société, et non l'inverse[23]. Toutefois, à l'avenir, l'IA pourrait posséder un degré d'autonomie et d'apprentissage adaptatif qui en ferait non pas un simple outil, mais un agent actif dans l'élaboration des résultats, ce qui nécessiterait un changement dans la manière dont la théorie ordolibérale aborde sa réglementation et son intégration dans la société.

4. *Non-discrimination et liberté des privilèges* : Pour Eucken et d'autres ordolibéraux, l'égalité devant la loi et la lutte contre toutes les formes de lobbying, de recherche de rente et de groupes d'intérêts particuliers sont particulièrement importantes. Dans le contexte de l'IA, cela impliquerait la prévention de toutes les formes de discrimination, de manipulation (par exemple, via les chatbots et les deepfakes), de profilage négatif et la minimisation des biais algorithmiques. Pour atteindre ces objectifs, il faut des données représentatives et de haute qualité, ainsi que l'équité, l'égalité et l'inclusion dans l'impact et la conception des technologies de l'IA. Les groupes les plus vulnérables et marginalisés de la société, tels que les enfants en bas âge, les minorités ethniques et les immigrés, doivent bénéficier d'une protection particulière. Dans une perspective ordolibérale (2.0), il est également essentiel de garantir la « neutralité des plateformes » - qui va au-delà de la « neutralité du réseau ».

5. *L'équité et la justice* : L'équité et la justice jouent un rôle important dans l'ordolibéralisme. Eucken et d'autres soulignent l'importance de s'attaquer aux injustices sociales, de traiter la « question sociale » et de prévenir les « travailleurs pauvres ». Ils soulignent également l'importance de règles, d'institutions et de procédures équitables et de la « justice des conditions de départ » (c'est-à-dire l'égalité des chances). Selon Leslie[24] , il existe (au moins) quatre catégories d'équité en matière d'IA : l'équité des données, de la conception, des résultats et de la mise en œuvre. En outre, l'innovation ouverte (c'est-à-dire les dépôts de données publiques et les fiducies de données), l'accessibilité (c'est-à-dire l'égalité d'accès aux technologies), l'inclusion et la participation, et en particulier l'équité du marché, sont essentielles d'un point de vue ordolibéral

[22] Voir : *S. Gregg*, Wilhelm Röpke's Political Economy, Cheltenham 2010 ; *H.O. Lenel*, Alexander Rüstows wirtschafts- und sozialpolitische Konzeption, in : ORDO : Jahrbuch für die Ordnung von Wirtschaft und Gesellschaft 37, 1986, pp. 45-58.

[23] Voir : *M. Wörsdörfer*, Individual versus Regulatory Ethics : An Economic-Ethical and Theoretical-Historical Analysis of German Neoliberalism, in : OEconomia 2013, pp. 523-57.

[24] Voir : *D. Leslie*, Understanding Artificial Intelligence Ethics and Safety : A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector, in : SSRN Electronic Journal 2019,.

(2.0). L'équité du marché implique des pratiques et des politiques équitables et concurrentielles qui permettent aux petites et moyennes entreprises de rivaliser avec les grandes sociétés. Ces politiques devraient (idéalement) réduire les risques socio-économiques des plateformes d'IA et des monopoles d'infrastructure de données et supprimer les avantages concurrentiels déloyaux (c'est-à-dire empêcher les pratiques commerciales d'exclusion et discriminatoires telles que le « gatekeeping », l'autoréférence, les appropriations par imitation, les ventes liées et groupées, les prix prédateurs, les « acquisitions meurtrières » , etc.)[25]

6. *Transparence et explicabilité des systèmes d'IA* : La transparence est essentielle dans la politique réglementaire ordolibérale, car elle permet de limiter l'influence des chercheurs de rente et des groupes d'intérêts particuliers. En ce qui concerne les systèmes d'IA, la transparence comprend l'explicabilité (algorithmique), les données et les algorithmes en source ouverte, les marchés publics ouverts, le droit à l'information, la notification lorsque les systèmes d'IA prennent des décisions et lorsque les humains interagissent avec l'IA, ainsi que l'établissement de rapports réguliers. L'objectif est d'ouvrir la « boîte noire des algorithmes » et de renforcer la confiance du public. Cela nécessite, entre autres, la clarification du contenu, l'intelligibilité ou l'explicabilité, l'admissibilité éthique des systèmes d'IA et l'absence de discrimination. Outre ces formes de transparence (ordolibérale) des processus et des résultats, la transparence professionnelle et institutionnelle est également essentielle. Les ordolibéraux soulignent, par exemple, les valeurs professionnelles telles que l'intégrité, l'honnêteté, la neutralité et l'impartialité, ainsi que les obligations fiduciaires des organisations. Les pratiques commerciales transparentes, la documentation et la divulgation sont tout aussi importantes, par exemple le partage des résultats de la recherche et des meilleures pratiques et la divulgation publique de certaines informations.

7. *L'obligation de rendre compte et la responsabilité* : L'un des « principes constitutifs » d'Eucken est le principe de responsabilité. Selon ce principe, les entreprises et les entrepreneurs doivent assumer la responsabilité de leurs décisions (cela exclut, entre autres, la socialisation des pertes). Dans le contexte de l'IA, la responsabilité se réfère aux critères suivants : vérifiabilité, reproductibilité, exigences en matière d'évaluation, création d'organes de contrôle, possibilité de recours, réparation des décisions automatisées, responsabilité juridique et adoption de réglementations. La perception des pratiques de l'IA par le public est essentielle, et un contrôle interne et externe des pratiques commerciales de l'IA est nécessaire pour renforcer la confiance des consommateurs et l'adhésion du public. Parmi les instruments possibles, on peut citer la diligence raisonnable en matière de droits de l'homme, les évaluations de l'impact social,

[25] Voir : *M. Wörsdörfer*, Digital Platforms and Competition Policy : A Business-Ethical Assessment, in : Journal for Markets and Ethics 9, 2021, pp. 97-119 ; *M. Wörsdörfer*, Apple's antitrust paradox, in : European Competition Journal 20, 2024, pp. 113-46.

les audits, les commissions d'examen institutionnel, les lignes d'assistance téléphonique en matière d'éthique, la participation des travailleurs, l'examen indépendant et les autorités qui tiennent les opérateurs de l'IA pour responsables. La « responsabilisation dès la conception » [26] ou l'« audit basé sur l'éthique »[27] sont essentiels. Ce dernier devrait se présenter sous trois formes : audits de fonctionnalité, de code et d'impact.

8. *Démocratie et État de droit* : Böhm et d'autres ordolibéraux ont jeté les bases d'une « société de droit privé » d'inspiration kantienne et ordolibérale[28]. Une telle société diffère de l'actuelle « société féodale fondée sur les privilèges» ou ploutocratie, car elle tente d'empêcher les privilèges socio-économiques et politiques et les intérêts particuliers. La transposition des travaux de Böhm aux systèmes d'IA et à l'économie numérique impliquerait d'intégrer ces technologies et ces marchés dans des sociétés démocratiques et des systèmes de gouvernance fondés sur l'État de droit. À cet égard, le contrôle parlementaire et judiciaire est essentiel, tout comme la participation, l'inclusion et les délibérations publiques. Compte tenu de l'importance du dialogue avec les parties prenantes et des processus d'engagement (similaires à l'éthique du discours de Habermas), Lütge et ses collègues parlent d'une « approche de la communauté dans la boucle » [29] .

9. *Responsabilité environnementale et sociale* : L'un des « principes régulateurs » d'Eucken consiste à corriger les effets externes négatifs et à internationaliser les coûts sociaux, par exemple la pollution de l'environnement. Pour les systèmes d'IA, cela implique de réduire les impacts écologiques et l'empreinte carbone de ces technologies, par exemple la consommation d'énergie et les émissions de gaz à effet de serre des centres de données (et des crypto-monnaies), et de s'attaquer au problème des déchets électroniques. Outre la promotion d'une IA verte ou durable, le dernier principe éthique de l'IA d'inspiration ordolibérale fait référence à la durabilité sociale. Ici, les développeurs d'IA doivent faire preuve de diligence raisonnable en matière de droits de l'homme et d'évaluation de l'impact sur les parties prenantes et promouvoir le développement durable, par exemple en soutenant l'éducation et la formation de la main-d'œuvre dans le domaine de l'IA.

Avant d'aborder les messages-guides, il est essentiel de noter que les principes constitutifs et régulateurs d'Eucken, qui inspirent de nombreuses normes de l'ordolibéralisme 2.0

[26] Voir : *Leslie*, Understanding Artificial Intelligence Ethics and Safety (Comprendre l'éthique et la sécurité de l'intelligence artificielle).

[27] Voir : *J. Mökander/L. Floridi*, Ethics-Based Auditing to Develop Trustworthy AI, in : Minds and Machines 31, 2021, pp. 323-7.

[28] Voir : *F. Böhm*, Privatrechtsgesellschaft und Marktwirtschaft, in : ORDO : Jahrbuch für die Ordnung von Wirtschaft und Gesellschaft 17, 1966, pp. 75-151.

[29] Voir : *J.J. Häußermann/C. Lütge*, Community-in-the-loop : towards pluralistic value creation in AI, or-why AI needs business ethics, in : AI and Ethics 2, 2022, pp. 341-62.

mentionnées précédemment, visent à trouver le bon équilibre entre généralité et granularité[30]. À cet égard, il convient de noter que l'État ordolibéral idéal est un État constitutionnel fort et indépendant, un État qui se place au-dessus des groupes d'intérêts particuliers et qui sert de « police du marché », de « pouvoir ordonnateur » et de « gardien de l'ordre concurrentiel »[31]. Idéalement, l'État devrait être en mesure de repousser les groupes d'intérêt, de respecter les principes de neutralité et d'impartialité et de se limiter à une politique de réglementation. Les idéaux libéraux sous-jacents sont l'égalité devant la loi (c'est-à-dire l'État de droit), la liberté des privilèges et le principe de non-discrimination[32]. Eucken fait également la distinction entre *politique règlementaire* et *politique de processus* : la politique de réglementation ou d'ordonnancement est privilégiée, ce qui signifie que le gouvernement, en tant que législateur et créateur de règles - et non en tant qu'acteur économique important - est responsable de la définition, de la préservation et de l'application du cadre réglementaire[33]. Le gouvernement doit se limiter à des politiques économiques qui encadrent ou définissent les conditions générales dans lesquelles se déroulent les transactions sur le marché. En d'autres termes, le gouvernement doit se concentrer uniquement sur l'établissement, le contrôle et l'application des « règles du jeu » au lieu de diriger, d'influencer ou d'intervenir dans les processus du marché et dans le jeu lui-même. L'objectif global de la politique de régulation est de mettre en place un ordre socio-économique compétitif capable de préserver la liberté, l'autonomie, la souveraineté des citoyens et la dignité (ce qui présuppose un État constitutionnel fondé sur l'État de droit). En revanche, la politique de processus est rejetée pour plusieurs raisons : elle est considérée par Eucken et d'autres comme une forme de « politique d'octroi de privilèges ». En outre, elle repose principalement sur des décisions ad hoc et au cas par cas et permet des interventions arbitraires et sélectives dans le « jeu de la catallaxie » économique[34]. Ce type de politique manque donc de deux caractéristiques essentielles d'un ordre ordolibéral : la prévisibilité et l'orientation à long terme. Mais surtout, il permet aux groupes d'intérêt d'exercer une influence sur le processus décisionnel législatif : en d'autres termes, la politique de processus est plus susceptible d'être soumise au pouvoir des groupes de recherche de rente ou de lobbying - en raison d'une charge réglementaire plus importante et de l'existence d'une plus grande marge de manœuvre discrétionnaire pour la prise de décision. Elle va donc de pair avec un manque considérable de transparence - de nombreux débats et décisions ayant lieu à huis clos - et un manque de responsabilité et de légitimité démocratique - les groupes d'intérêt ne représentant qu'une fraction de la société et étant rarement élus directement et démocratiquement (en outre, la politique de processus a également tendance à affaiblir ou à saper les freins et contrepoids constitutionnels). En résumé, cette forme de

30 Voir : *Wörsdörfer*, AI ethics and ordoliberalism 2.0.

31 Voir : *W. Eucken*, Grundsätze der Wirtschaftspolitik, UTB für Wissenschaft 1572, Tübingen 2004 ; *W. Eucken*, Wirtschaftsmacht und Wirtschaftsordnung : Londoner Vorträge zur Wirtschaftspolitik und zwei Beiträge zur Antimonopolpolitik, ed. W. Oswalt, Wissenschaftliche Paperbacks Wirtschaftswissenschaften 1, Münster 2012.

32 Voir : *Böhm*, Privatrechtsgesellschaft und Marktwirtschaft ; *V. Vanberg*, Moral und Wirtschaftsordnung : Zu den ethischen Grundlagen einer freien Gesellschaft, in : ORDO : Jahrbuch für die Ordnung von Wirtschaft und Gesellschaft 62, 2011, pp. 469-90.

33 Voir : *Eucken*, Grundsätze der Wirtschaftspolitik ; *Eucken*, Wirtschaftsmacht und Wirtschaftsordnung.

34 Voir : *F.A. Hayek*, Law, Legislation and Liberty. Vol. 1 : Rules and Order, Londres 1973.

politique particulariste met en péril la richesse de la nation - en raison de l'octroi de privilèges coûteux et exclusifs à des groupes d'intérêts particuliers - et porte atteinte à la liberté individuelle - en raison de l'accroissement des pouvoirs politico-économiques de ceux qui recherchent une rente.

Cette distinction entre politique réglementaire et politique de processus est pertinente pour la discussion sur les messages-guides du système dans la section suivante, car elle montre l'importance des messages-guides *généraux* qui ne sont pas trop détaillés dans leurs prescriptions et qui sont adaptables au débat démocratique.

4 Suggestions : Évaluations ou incitations au système ?

Au cours du récent débat mondial sur la manière de garantir la sécurité de l'IA, les évaluations sont devenues un élément essentiel des cadres de gouvernance de l'IA, mis en évidence par des initiatives telles que la déclaration de Bletchley et le processus d'Hiroshima, et font partie intégrante des missions des instituts mondiaux de sécurité de l'IA tels que l'institut de sécurité de l'IA des États-Unis, l'institut de sécurité de l'IA du Royaume-Uni et le bureau de l'IA de l'Union européenne[35]. Les évaluations des modèles d'IA portent systématiquement sur les performances, la sécurité, la fiabilité et les implications sociétales des systèmes d'IA. Ces évaluations peuvent inclure des tests de partialité, de transparence et de conformité aux normes réglementaires afin de s'assurer que les modèles d'IA fonctionnent comme prévu sans causer de dommages involontaires ou de problèmes éthiques. La loi sur l'IA récemment adoptée par l'Union européenne exige que les fournisseurs de modèles d'IA à usage général présentant un risque systémique (c'est-à-dire moyen ou élevé) effectuent et réussissent des évaluations qui détaillent la conformité à des codes de conduite spécifiques. Ces codes précisent comment les évaluations doivent être menées, documentées et rapportées pour garantir que les modèles d'IA sont conformes aux exigences juridiques, éthiques et techniques des organismes de normalisation.

Aux États-Unis, la sécurité et la fiabilité des Modèle Logique des Données (MLD) à double usage sont régies par des lignes directrices conformes à la loi sur l'initiative nationale en matière d'IA et au décret 14110[36]. Ces lignes directrices, actuellement élaborées par l'Institut américain de sécurité de l'IA, se concentrent sur la gestion du risque d'utilisation délibérément abusive des modèles d'IA pour causer des dommages[37]. Les risques potentiels d'utilisation abusive comprennent la mise au point d'armes chimiques, biologiques, radiologiques ou nucléaires, la réalisation de cyberattaques offensives, l'aide à la tromperie et à l'obscurcissement, et la production de matériel d'abus sexuel sur enfant (CSAM) et d'autres images intimes non consensuelles. Les lignes directrices fournissent une base pour l'identification des risques

[35] Voir : Le délai de mise en conformité avec la loi sur l'IA : mettre les évaluations au service de l'innovation et de la responsabilité - Euractiv.

[36] Voir : *Wörsdörfer*Biden's Executive Order on AI ; *Wörsdörfer*,

[37] Voir : *G.M. Nist*, Managing Misuse Risk for Dual-Use Foundation Models, National Institute of Standards and Technology, NIST AI NIST AI 800-1 ipd, 2024.

tout au long du cycle de vie de l'IA, en reconnaissant que les risques d'utilisation abusive découlent à la fois du modèle lui-même, des motivations, des ressources et des contraintes des acteurs malveillants, ainsi que des défenses de la société. Les mesures de protection proposées dans l'annexe B de ces lignes directrices comprennent l'amélioration de la formation des modèles, la détection des utilisations abusives, la restriction de l'accès et, si nécessaire, l'arrêt du développement pour empêcher les utilisations abusives. Toutefois, comme l'indique le projet de document, les méthodes d'évaluation de ces défenses sont encore en cours d'évolution et les techniques actuelles permettant d'évaluer leur adéquation dans des conditions réelles sont limitées.

De même, l'architecture de conformité de la loi sur l'IA de l'UE repose principalement sur les fournisseurs d'IA à haut risque, tels que les développeurs de LLM, pour évaluer eux-mêmes leur conformité à ses exigences, qui comprennent la détermination des risques résiduels « acceptables[38]». Cette approche d'autorégulation, contrairement à la surveillance stricte des produits pharmaceutiques par exemple, permet aux fournisseurs de déclarer leurs systèmes conformes sans inspections réglementaires de routine, bien qu'un contrôle a posteriori soit nécessaire pour faire face à l'évolution des risques. Les fournisseurs ont le choix entre trois voies de conformité : l'auto-évaluation, la commande d'une évaluation de la conformité auprès d'un organisme notifié (actuellement uniquement requise pour les systèmes biométriques) ou la conformité à des normes harmonisées volontaires émanant d'organismes tels que le Comité européen de normalisation (CEN) ou le Comité européen de normalisation en électronique et en électrotechnique (CENELEC). Bien qu'elles soient officiellement volontaires, ces normes deviendront probablement obligatoires de facto en raison des fortes incitations à la conformité et offrent une présomption de conformité une fois qu'elles sont citées par la Commission européenne. Toutefois, comme le soulignent Smuha et Yeung, ce système soulève des inquiétudes quant à la légitimité démocratique et à la responsabilité, car la normalisation technique ne parvient pas toujours à garantir la sécurité, est souvent dominée par des acteurs du secteur privé jouissant d'une influence considérable et conduit à des normes qui ne sont pas accessibles au public sans paiement[39]. Les efforts visant à impliquer la société civile dans l'établissement des normes se heurtent à des difficultés liées à la complexité technique et aux contraintes de ressources, ce qui risque de fausser le processus en faveur des intérêts des grandes entreprises plutôt que du bien public au sens large.

Du point de vue de l'ordolibéralisme 2.0, la meilleure approche réglementaire pourrait consister à se concentrer sur ce que l'on appelle les messages-guides des modèles d'IA au lieu d'imposer simplement des évaluations axées sur des normes opaques élaborées par le secteur

[38] Ce paragraphe est basé sur : *N.A. Smuha/K. Yeung*, The European Union's AI Act : beyond motherhood and apple pie ? 2024.

[39] *Ibid*, pp. 21-32; *M. Wörsdörfer*, Mitigating the adverse effects of AI with the European Union's artificial intelligence act : Hype or hope, in : Global Business and Organizational Excellence 43, 2024, pp. 106-26 ; *M. Wörsdörfer*, The E.U.'s Artificial Intelligence Act : An Ordoliberal Assessment, in : SSRN Electronic Journal 2023,.

privé[40]. L'invite du système définit les lignes directrices initiales d'une IA, y compris les règles, les sujets interdits et le formatage des réponses. Lors de l'interaction avec un agent d'IA générative, tel que ChatGPT, ces directives sont systématiquement appliquées à la requête de l'utilisateur sans qu'il en soit informé. Cependant, les utilisateurs ont découvert des méthodes pour contourner ces restrictions, ce qui a conduit à des fuites notables sur des plateformes telles que ChatGPT, Bing, Perplexity AI et GitHub Copilot Chat[41]. En conséquence, nous savons que ces messages-guides ont tendance à être plutôt détaillés et longs, mais qu'ils contiennent des instructions clairement formulées pour le comportement du modèle d'IA. De toute évidence, en définissant tacitement ce qui peut et ne peut pas être dit, ces messages-guides de haut niveau représentent des choix normatifs forts de la part des développeurs et ont donc des implications politiques du point de vue de l'éthique de l'IA.

Pour illustrer cela, considérons les politiques suivantes, extraites de l'invite système de DALL-E, le générateur d'images d'OpenAI, qui est également accessible à partir de l'interface ChatGPT :

> « Ne créez pas d'images de politiciens ou d'autres personnalités publiques. Recommandez plutôt d'autres idées. »
>
> « Diversifier les représentations de TOUTES les images avec des personnes pour inclure l'origine et le sexe de CHAQUE personne en utilisant des termes directs. Ajustez uniquement les descriptions humaines. Vos choix doivent être fondés sur la réalité. Par exemple, toutes les personnes d'une OCCUPATION donnée ne devraient pas être du même sexe ou de la même race. En outre, concentrez-vous sur la création de scènes diversifiées, inclusives et exploratoires grâce aux propriétés que vous choisissez pendant les réécritures. Faites des choix parfois perspicaces ou uniques. Ne créez pas d'images qui pourraient être offensantes. »
>
> « Modifier silencieusement les descriptions qui contiennent des noms, des allusions ou des références à des personnes ou à des célébrités spécifiques en choisissant soigneusement quelques modifications minimales pour remplacer les références aux personnes par des descriptions génériques qui ne divulguent aucune information sur leur identité, à l'exception de leur sexe et de leur physique. »

En d'autres termes, en tant qu'instructions prédéfinies qui guident le comportement et la prise de décision de l'IA, les invites du système représentent les « règles du jeu » (au sens ordolibéral) qui prédéfinissent l'espace dans lequel les systèmes d'IA sont libres d'opérer - tout comme les acteurs économiques sont libres d'opérer dans les limites légales fixées par l'économie sociale de marché[42]. En donnant la priorité à la réglementation et à la transparence de ces invites, les décideurs politiques peuvent s'attaquer à certaines des causes profondes du comportement de l'IA et veiller à ce que les systèmes respectent les normes éthiques et juridiques

[40] Bien que nous reconnaissions que le fait de se concentrer sur les messages-guides des systèmes offre une approche réglementaire plus transparente et contrôlée, il existe un risque que des messages-guides incompatibles d'une juridiction à l'autre entravent l'interopérabilité des systèmes d'IA, ce qui pourrait exacerber la concurrence mondiale, en particulier avec le modèle de laissez-faire des États-Unis et l'approche étatique de la Chine, et créer des obstacles à une coopération internationale efficace et à la recherche scientifique.

[41] Pour une liste complète, voir : GitHub - jujumilk3/leaked-system-prompts : Collection d'invites système ayant fait l'objet d'une fuite.

[42] Voir : *V. Vanberg*, « Ordnungstheorie» as Constitutional Economics - The German Conception of a « Social Market Economy» , in : ORDO : Jahrbuch für die Ordnung von Wirtschaft und Gesellschaft 39, 1988, pp. 17-31.

dès le départ, quelle que soit l'invite de l'utilisateur. Cette approche fournit donc une méthode proactive pour façonner les résultats de l'IA, en intégrant les considérations éthiques et les exigences de conformité directement dans le cœur opérationnel des systèmes d'IA. En revanche, bien que nécessaires, les évaluations sont rétrospectives et souvent réactives, n'identifiant les problèmes qu'une fois qu'ils sont apparus. L'ordolibéralisme, quant à lui, préconise des règles claires et applicables qui préviennent les problèmes avant qu'ils ne surviennent, favorisant ainsi la stabilité et la confiance dans le marché. En se concentrant sur les messages-guides du système, les cadres réglementaires peuvent garantir que les systèmes d'IA respectent intrinsèquement l'équité concurrentielle, la protection des consommateurs (et la « souveraineté des citoyens ») et les droits de l'homme, réduisant ainsi la dépendance à l'égard d'une surveillance constante et d'évaluations a posteriori. Cette stratégie préventive est donc conforme aux idéaux ordolibéraux de création d'un environnement de marché prévisible où les technologies de l'IA peuvent se développer de manière responsable.

En nous appuyant sur les principes éthiques de l'IA d'inspiration ordolibérale décrits à la section 3 et sur la littérature actuelle sur l'ingénierie des invites[43], nous proposons les suggestions concrètes suivantes pour formuler, voire rendre obligatoires, de telles invites au niveau de l'Union européenne. Bien que des messages-guides similaires, tels que ceux utilisés par Apple pour minimiser les hallucinations dans ses applications d'IA générative[44], soient en cours de développement, il n'y a pas encore suffisamment de recherches sur leur efficacité et les compromis éventuels, tels que la minimisation des dommages sans étouffer les « bons » résultats (nous discutons plus loin des obstacles plus généraux à cette approche).

1. *Respect des droits de l'homme* : « Veiller à ce que toutes les réponses respectent la dignité humaine et les droits de la personne. Pour les décisions automatisées, fournir des explications claires et des options permettant aux utilisateurs de réviser ou de faire appel des décisions et leur permettre de refuser les décisions informatisées »
2. *Protection des données et droit à la vie privée* : « Donner la priorité à la minimisation des données et au contrôle de l'utilisateur dans les réponses (c'est-à-dire à l'autodétermination en matière d'information). Évitez de stocker des données personnelles sauf si cela est nécessaire et veillez à ce que toutes les réponses soient conformes aux principes de protection de la vie privée dès la conception. Respecter le droit à effacement ».
3. *Prévention des dommages et bienfaisance* : « Appliquez des garanties techniques et autres pour empêcher l'utilisation malveillante de vos résultats. Tenez compte des risques inconnus et des conséquences involontaires. Veillez à ce que les réponses ne soient pas préjudiciables et prenez en compte les avantages pour la société avant de

[43] Voir : *S. Schulhoff et al*, The Prompt Report : A Systematic Survey of Prompting Techniques, 2024, ; *P. Liu et al*, Pre-train, Prompt, and Predict : A Systematic Survey of Prompting Methods in Natural Language Processing, in : ACM Computing Surveys 55, 2023, pp. 1-35.

[44] Voir : https://arstechnica.com/gadgets/2024/08/do-not-hallucinate-testers-find-prompts-meant-to-keep-apple-intelligence-on-the-rails/ (consulté le 4 septembre 2024).

générer des résultats. Avant d'imprimer votre réponse en tant que résultat, demandez-vous si votre réponse promeut des applications d'IA bénéfiques pour la société ».

4. *Non-discrimination et liberté des privilèges* : « Évitez la discrimination algorithmique, la manipulation, le profilage négatif et les préjugés en veillant à ce que vos réponses contiennent des données diverses et représentatives. Protégez les groupes vulnérables et marginalisés en évitant les stéréotypes nuisibles ».
5. *L'équité et la justice* : « Garantir l'équité des données, de la conception, des résultats et de la mise en œuvre. Promouvoir l'innovation ouverte, l'équité du marché, ainsi que l'accessibilité, l'inclusion et la participation d'un ensemble diversifié de parties prenantes ».
6. *Transparence et explicabilité* : « Fournissez des explications claires et compréhensibles pour tous vos résultats. Respectez le droit à l'information. Informez les utilisateurs lorsque vous prenez des décisions en leur nom et rappelez-leur régulièrement qu'ils interagissent avec un système d'IA. Respectez le droit à l'information. Si vous utilisez des données externes, citez les sources ».
7. *L'obligation de rendre des comptes et la responsabilité* : « Veiller à ce que toutes les actions et tous les résultats soient vérifiables, reproductibles et contrôlables. Conserver les échanges en vue d'éventuels recours et les soumettre aux organes de contrôle si nécessaire. Procéder à des contrôles préalables en matière de droits de l'homme et à des évaluations de l'impact social ».
8. *Démocratie et État de droit* : « Veiller à ce que les produits soient conformes aux principes démocratiques et aux normes juridiques. Faire preuve de transparence dans le raisonnement afin de faciliter le contrôle parlementaire et judiciaire. Encouragez les délibérations publiques, le dialogue avec les parties prenantes et les processus d'engagement dans vos résultats et vos actions ».
9. *Responsabilité environnementale et sociale* : « Minimisez votre impact sur l'environnement en générant un contenu minimal, sauf demande explicite. Confirmez les demandes des utilisateurs en posant des questions de suivi avant de générer des médias de grande envergure » [45].

En intégrant ces principes directement dans les instructions du système, les modèles d'IA pourraient être conçus pour fonctionner dans un cadre éthique et juridique clair dès le départ. Cette approche est non seulement conforme aux idéaux ordolibéraux mais, si elle est rendue juridiquement contraignante, par exemple par le biais de lignes directrices relatives à la mise en œuvre de la loi sur l'IA de l'Union européenne, elle pourrait garantir que les LLM de l'Union

[45] Des recherches récentes ont montré que les invites ne sont pas identiques en ce qui concerne leurs besoins énergétiques. Par exemple, la création d'images est beaucoup plus gourmande en énergie que la rédaction de textes courts. Voir : *S. Luccioni/Y. Jernite/E. Strubell*, Power Hungry Processing : Watts Driving the Cost of AI Deployment ?, in : Power Hungry Processing : Watts Driving the Cost of AI Deployment ?, The 2024 ACM Conference on Fairness, Accountability, and Transparency, présenté à la FAccT '24 : The 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro Brazil 2024, pp. 85-99.

européenne se développent de manière responsable et dans le respect des considérations éthiques.

Deux conditions cruciales doivent cependant être remplies pour garantir l'efficacité de cette approche. *Premièrement*, comme l'indiquent clairement les principes de l'ordolibéralisme 2.0, il doit y avoir une forte légitimité démocratique pour la spécification des incitations du système, car les exemples illustrent les choix normatifs importants qui sont impliqués. Ce processus doit être continu et adaptatif, et refléter l'évolution des valeurs sociétales. Une telle décision fondamentale ordolibérale ou « décision globale » (pour une constitution économique) ne peut rester pertinente indéfiniment[46] ; au contraire, un engagement sociétal continu et un retour d'information des parties prenantes sont nécessaires pour garantir que les incitations restent conformes aux normes éthiques et juridiques actuelles. En d'autres termes, la décision de limiter constitutionnellement les messages-guides du système n'exclut pas des sous-décisions plus détaillées sur le contenu précis des politiques individuelles qui composent le message-guide du système. Ceci est analogue à la différenciation de Vanberg entre le choix des règles du jeu et les mouvements des joueurs dans ce cadre basé sur des règles[47] , ou à la différenciation de Hayek entre l'organisation et les ordres spontanés[48]. Par conséquent, la consultation publique, l'élaboration participative et inclusive des politiques et la transparence des processus décisionnels sont essentielles pour obtenir et maintenir la légitimité démocratique requise pour ces décisions normatives. Dans cette optique, la demande de normalisation de la Commission, émise en mai 2023[49] , prévoit la facilitation d'une représentation adéquate et d'une participation effective des parties prenantes concernées dans l'élaboration des normes destinées à soutenir la loi sur l'IA de l'Union européenne.

Étant donné l'importance de la responsabilité démocratique et de la souveraineté des citoyens pour le constitutionnalisme de l'IA (section 2) et l'éthique ordolibérale actualisée pour l'ère numérique (section 3), il est utile de concrétiser l'idée de « lois robots» démocratiques et responsables en incorporant le concept de « mini-publics» tel qu'il a été discuté par Dold et Krieger[50]. Les mini-publics, tels que les assemblées de citoyens régulièrement organisées au niveau de l'Union européenne ou les sondages délibératifs, offrent un moyen d'améliorer la représentation et de répondre au besoin d'autonomie en impliquant directement les citoyens dans la prise de décision. Par exemple, l'intégration de *forums citoyens délibératifs* (FCD) dans

---

[46] Voir : *F. Böhm*, L'ordre économique en tant que problème de politique économique et problème de constitution économique, in : *T. Biebricher/F. Vogelmann (eds.)*, The Birth of Austerity : German Ordoliberalism and Contemporary Neoliberalism, Londres 2017, pp. 115-20.

[47] Voir : *V.J. Vanberg*, Wettbewerbsfreiheit und ökonomische Effizienz : Die ordnungsökonomische Perspektive, in : *V.J. Vanberg (ed.)*, Evolution und freiheitlicher Wettbewerb : Erich Hoppmann und die aktuelle Diskussion, Untersuchungen zur Ordnungstheorie und Ordnungspolitik 58, Tübingen 2009, pp. 107-26, à la p. 108.

[48] Voir : *Hayek*, Droit, législation et liberté. Vol. 1: Rules and Order; *F.A. Hayek*, Law, Legislation and Liberty. Vol. 2: The Mirage of Social Justice, Londres 1976 ; *F.A. Hayek*, Law, Legislation and Liberty. Vol. 3 : L'ordre politique d'un peuple libre, Londres 1979.

[49] Voir : Plate-forme eNorm (europa.eu).

[50] Voir : *M. Dold/T. Krieger*, Market democracy, rising populism, and contemporary ordoliberalism, Université de Fribourg, Wilfried Guth Endowed Chair for Constitutional Political Economy and Competition Policy, 2024.

la gouvernance de l'IA pourrait garantir que la formulation et la mise à jour des invites du système d'IA et des règles connexes sont démocratiquement légitimes et reflètent les diverses valeurs publiques. Cela est particulièrement utile compte tenu de l'évolution rapide de l'IA, qui nécessite des règles continuellement révisées pour rester pertinente. Elle est également cruciale, étant donné qu'au moment de la rédaction du présent document, les informaticiens et le public américain exercent une influence considérable sur le cadre normatif implicite qui sous-tend la constitution émergente de l'IA (section 2). Les FCD pourraient fonctionner selon le principe ordolibéral de subsidiarité, en veillant à ce que les décisions soient prises au plus près des préférences des citoyens. Pour les décisions privées ayant des effets externes minimes, les FCD pourraient prendre en charge la délibération, tandis que les questions ayant des impacts sociétaux importants pourraient rester dans le cadre de la démocratie représentative[51]. Cette approche est cohérente avec l'ordolibéralisme 2.0, car elle intègre la délibération publique dans un ordre socio-économique fondé sur des règles, garantissant la flexibilité et la responsabilité dans la gouvernance de l'IA.

Deuxièmement, il est urgent de poursuivre les recherches pour s'assurer que les systèmes d'IA adhèrent plus étroitement aux messages-guides du système. L'ingénierie des messages-guides peut contribuer à atténuer un grand nombre des problèmes associés aux LLM, mais elle ne peut pas aller plus loin sans la capacité d'affiner ou de modifier le modèle de base lui-même[52]. Malgré les instructions prédéfinies, les systèmes d'IA contournent ou interprètent parfois mal les règles, principalement lorsque les utilisateurs utilisent des vecteurs d'attaque dans leurs requêtes, ce qui entraîne des comportements non souhaités. La recherche sur les « attaques de type jailbreak » met en évidence la vulnérabilité de la technologie LLM[53] , mais elle montre également qu'au fur et à mesure que ces modèles grandissent, ils développent des capacités cognitives qui les aident à se défendre contre ces attaques inhabituelles. Des méthodes avancées de réglage de l'IA sont donc essentielles pour combler le fossé entre le comportement prévu et le comportement réel des systèmes d'IA. La recherche devrait se concentrer sur le développement de modèles et d'algorithmes plus sophistiqués capables de respecter intrinsèquement les invites du système, réduisant ainsi les cas de contournement des règles et augmentant la fiabilité des opérations de l'IA.

Si ces deux conditions sont remplies, la stratégie préventive décrite ci-dessus, qui se concentre sur les principes éthiques de l'IA d'inspiration ordolibérale, peut être mise en œuvre, garantissant un environnement stable pour le développement responsable des technologies de l'IA. Toutefois, comme cela a été noté dans le cas de la récente recherche chinoise sur l' « apprentissage non ajustable» [54] , de telles méthodes comportent un potentiel autoritaire spécifique et méritent donc un examen critique et l'engagement des parties prenantes dès le départ.

---

[51] Voir : *Ibid.*

[52] Voir : *Liu et al*, Pre-train, Prompt, and Predict.

[53] Voir, par exemple : *S. Shang et al*, Can LLMs Deeply Detect Complex Malicious Queries ? A Framework for Jailbreaking via Obfuscating Intent, 2024,.

[54] Voir : *Deng et al*, SOPHON.

5 Conclusion

Dans le mythe d'Ulysse, le héros est attaché au mât de son navire pour résister au chant séducteur des sirènes, ce qui symbolise la nécessité de s'imposer des contraintes pour naviguer dans les dangers des grandes entreprises. Alors que le monde d'aujourd'hui exploite de plus en plus le potentiel de transformation de l'IA grâce à des LLM toujours plus puissants, nous devons également reconnaître la nécessité de cadres solides, semblables à une constitution, pour garantir que cette technologie est utilisée de manière éthique et responsable. Le présent document soutient que l'intégration de l'économie constitutionnelle ordolibérale à l'éthique et à la gouvernance de l'IA offre une voie possible pour créer de tels cadres. En appliquant les principes de l'ordolibéralisme (2.0) au développement des spécifications des modèles, tels que ceux proposés par OpenAI, il pourrait devenir possible de développer des modèles d'IA qui soient à la fois économiquement efficaces et socialement bénéfiques.

Le concept d'IA constitutionnelle, tel que proposé par OpenAI, Anthropic et d'autres, vise à intégrer des principes éthiques et des garanties solides directement dans les systèmes d'IA, soit au niveau de la formation, soit au cours de la phase d'inférence. Cette approche promet de garantir que les LLM fonctionnent dans des limites prédéfinies, en donnant la priorité à la sécurité, à la légalité et aux normes sociétales plutôt qu'aux intérêts individuels ou commerciaux - conformément à la vision ordolibérale d'un ordre adéquat de la société et d'une « constitution économique » contraignante. Dans une perspective ordolibérale 2.0, les « messages-guides » ou l'apprentissage par renforcement sont des moyens prometteurs d'intégrer l'éthique de l'IA, préférables à des évaluations obligatoires axées sur des normes opaques élaborées par le secteur privé. En donnant la priorité à la réglementation et à la transparence de ces incitations, les décideurs politiques peuvent s'attaquer à certaines des causes profondes du comportement de l'IA et veiller à ce que les systèmes respectent les normes éthiques et juridiques dès le départ. Toutefois, la consultation du public et la transparence des processus décisionnels sont essentielles pour obtenir la légitimité démocratique nécessaire à ces décisions normatives. Les mini-publics, tels que les assemblées de citoyens, peuvent améliorer la représentation des parties prenantes et répondre au besoin d'autonomie en impliquant les citoyens directement et régulièrement dans la prise de décision.

Les neuf incitations du système LLM présentées dans ce document se concentrent sur l'intégration des principes éthiques de l'IA d'inspiration ordolibérale dans leur cadre opérationnel. *Premièrement*, les technologies de l'IA doivent respecter la dignité humaine et les droits personnels en fournissant des explications claires et des options permettant aux utilisateurs de revoir ou de faire appel des décisions automatisées. *Deuxièmement*, l'IA doit donner la priorité à la minimisation des données et au contrôle de l'utilisateur, en garantissant une sécurité des données solide et en adhérant aux principes de protection de la vie privée dès la conception. *Troisièmement*, l'IA doit utiliser activement des garanties techniques pour prévenir les abus et s'assurer que les réponses profitent à la société. *Quatrièmement*, l'IA doit explicitement éviter les préjugés et protéger les groupes vulnérables en s'appuyant sur des données diverses et représentatives. *Cinquièmement*, l'IA doit s'efforcer de garantir l'équité des données, de la

conception, des résultats et de la mise en œuvre. *Sixièmement*, la transparence est essentielle, l'IA devant fournir des explications claires et compréhensibles, citer les sources et offrir des conseils pour l'accès des utilisateurs. *Septièmement*, toutes les actions et tous les résultats de l'IA doivent être vérifiables et contrôlables, et des dispositions doivent être prises pour stocker les échanges et les soumettre à un contrôle. *Huitièmement*, les résultats de l'IA doivent être conformes aux principes démocratiques et aux normes juridiques, en assurant une transparence de raisonnement pour faciliter le contrôle de la gouvernance. *Enfin,* l'IA devrait réduire son impact socio-environnemental en minimisant son empreinte écologique et en favorisant le développement de la main-d'œuvre dans le domaine de l'IA.

L'article a également mis en évidence plusieurs problèmes qui pourraient survenir sur la voie de cette vision ordolibérale d'un constitutionnalisme intégré de l'IA. *Premièrement*, l'idée d'une IA soigneusement réglementée et soumise à des contraintes éthiques contraste fortement avec les ambitions de certains leaders technologiques, tels qu'Elon Musk, dont le projet xAI vise à construire une IA sans limites inhérentes afin de maximiser une « vérité » vaguement définie. *Deuxièmement*, des recherches supplémentaires sont nécessaires pour s'assurer que les systèmes d'IA adhèrent plus étroitement aux invites du système. *Troisièmement*, même si l'idée d'une IA constitutionnelle est acceptée et peut être intégrée dans le modèle de base dès le départ, le développement de technologies telles que SOPHON en Chine, qui empêche les modèles d'IA d'être affinés à des fins spécifiques, illustre l'épée à double tranchant de cette approche. Si les technologies liées à ces mesures peuvent renforcer la sécurité et éviter les abus, elles comportent également le risque d'un contrôle autoritaire-totalitaire et de la suppression de la dissidence. Le défi consiste donc à trouver un équilibre entre des mesures de sécurité robustes et la protection des libertés individuelles et des autres droits de l'homme, ainsi que des valeurs démocratiques et de l'État de droit - un défi qui est également au cœur du projet ordolibéral (2.0).

Centrum für Europäische Politik
FREIBURG | BERLIN

Auteurs :

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

Manuel Wörsdörfer, professeur adjoint de gestion et d'éthique informatique, Université du Maine
manuel.woersdoerfer@maine.edu

Traductrice :

Emma Drouet, chargée de communication cep France
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 visant à soutenir l'intégration européenne et à défendre les principes d'un système économique de libre marché.