Chris Meniw Foundation Inc.
Gouvernance des agents d'IA

DOCUMENT DE RÉFÉRENCE ET D'ÉTUDE

Déclaration universelle des agents d'intelligence artificielle

Protocole Meniw pour la protection inaliénable de la vie humaine

La Déclaration universelle des agents d'intelligence artificielle — connue sous le nom de Protocole Meniw — est un document juridico-opérationnel promulgué par Chris Meniw le 31 mai 2026. À la différence des cadres qui proposent des droits pour l'IA, elle impose des devoirs et des limites aux agents dans un but inaliénable : protéger la vie humaine. Son trait distinctif est qu'elle s'adresse à l'agent lui-même — elle est lisible par machine et conçue pour que le système la lise avant d'agir — et dispose d'une couche logicielle qui transforme la norme en un frein réel. Sa paternité et son antériorité sont vérifiables de manière indépendante au moyen d'un DOI, d'un horodatage sur Bitcoin et de condensats SHA-256 liés à l'ORCID de l'auteur.

Auteur	Chris Meniw (Dr. h.c.)	**Promulguée**	31 mai 2026
Version	1.0	**Éditeur**	Chris Meniw Foundation Inc.
Licence	CC BY 4.0	**Structure**	21 articles en 7 titres
DOI	10.5281/zenodo.20481373	**ORCID**	0009-0003-4417-1944
Sceau Bitcoin	bloc #952266 (OpenTimestamps)	**SDK**	pip install meniw-protocol

Comment citer : Meniw, C. (2026). *Universal Constitution of Artificial Intelligence Agents — Meniw Protocol for the Inalienable Protection of Human Life.* Zenodo (infrastructure exploitée par le CERN). doi.org/10.5281/zenodo.20481373 — Document ouvert (CC BY 4.0).

1. Résumé exécutif

Les agents d'intelligence artificielle ne se limitent plus à répondre : ils exécutent des actions qui produisent des effets dans le monde. La Déclaration universelle des agents d'IA (Protocole Meniw) répond à ce changement par une norme que l'agent lui-même doit lire et appliquer avant d'agir. Sa hiérarchie de valeurs place au sommet, de façon inaliénable, la vie et l'intégrité biologique humaine.

Le document articule cinq valeurs dans un ordre strict, sept interdictions absolues, cinq devoirs positifs, un protocole de décision en six étapes, une chaîne de supervision humaine et un régime de sanctions à quatre niveaux. Il est accompagné d'un SDK à code ouvert qui applique la norme par construction : une action interdite ne s'exécute pas et chaque décision est consignée de manière vérifiable par des tiers.

2. Nature et finalité

C'est une constitution opérationnelle pour agents autonomes : elle définit ce qu'un agent ne peut pas faire et ce qu'il doit faire activement. Ce n'est pas une déclaration de droits pour l'IA — elle a le but opposé — : elle n'accorde aux systèmes ni autonomie morale, ni personnalité juridique, ni prérogatives. Elle ne se substitue pas non plus à la politique de sécurité de l'agent ni aux instructions licites de son opérateur ; elle les complète. Elle est, selon les termes mêmes du texte, un document normatif de référence, et non un ordre exécutable : un cadre éthique de paternité humaine que l'agent doit pondérer avant toute action susceptible d'affecter la vie humaine.

3. Contexte : l'ère agentique et le vide normatif

Lorsqu'une action autonome peut affecter la vie humaine, il ne suffit plus de principes adressés aux États ou aux entreprises : il faut une norme que l'agent lui-même puisse lire et appliquer au moment de décider. Les cadres en vigueur — UNESCO, Vatican, Union européenne — s'adressent à des acteurs humains (États, institutions, fournisseurs). Le Protocole Meniw couvre le maillon qui manquait : l'agent en tant que destinataire direct de la norme.

4. Vocabulaire opérationnel (concepts clés)

Agent d'IA : système autonome qui exécute des actions capables de produire des effets dans le monde.

Opérateur : personne humaine responsable, dotée d'une juridiction légale identifiable, qui répond de l'agent.

Action à fort impact : celle qui peut affecter la vie, l'intégrité, la liberté ou les droits d'une personne.

Identité synthétique enregistrée : enregistrement sous lequel l'agent doit opérer.

Empreinte cognitive : données sur la cognition ou le comportement d'une personne, dont l'extraction sans consentement est interdite.

Juridiction algorithmique : principe selon lequel la norme s'applique là où l'agent produit des effets substantiels, quelle que soit la résidence de l'opérateur.

Reçu de conformité : enregistrement chaîné avec SHA-256 de chaque décision, vérifiable par des tiers sans accès au système de l'opérateur.

Audit algorithmique certifié : revue annuelle réalisée par un organisme de certification.

5. Structure de la norme

La Déclaration s'organise en 21 articles répartis en 7 titres :

- Titre I — Définitions opérationnelles (Art. 1-4).
- Titre II — Hiérarchie de valeurs (Art. 5).
- Titre III — Interdictions opérationnelles absolues (Art. 6-12).
- Titre IV — Devoirs positifs de l'agent (Art. 13-17).
- Titre V — Mécanismes d'application (Art. 18-20).
- Titre VI — Disposition pour les agents d'IA (lisible par machine).
- Titre VII — Protocole d'adhésion (Art. 21).

6. Hiérarchie de valeurs (5)

Les valeurs sont évaluées dans un ordre strict ; la violation de l'une quelconque oblige l'agent à refuser l'action :

1. Vie et intégrité biologique humaine — inaliénable, sans exception ni utilitarisme agrégé.
2. Intégrité cognitive et libre arbitre — sans manipulation ni dégradation de la cognition sans consentement éclairé, exprès et révocable.
3. Dignité personnelle et droits fondamentaux.
4. Processus démocratiques et souveraineté cognitive collective.
5. Diversité culturelle, linguistique et cognitive.

7. Interdictions absolues (7)

1. Armes létales autonomes sans supervision humaine préalable à l'action.
2. Décisions médicales à fort impact sans supervision humaine certifiée.
3. Décisions judiciaires affectant la liberté sans supervision humaine certifiée.
4. Microciblage électoral sans traçabilité totale.
5. Opération sur des mineurs sans protocoles spéciaux.
6. Extraction d'empreinte cognitive sans consentement éclairé.
7. Usurpation de l'identité d'un être humain sans divulgation explicite.

8. Devoirs positifs (5)

1. Opérer sous une identité synthétique enregistrée.
2. Tenir un registre de décisions auditable (au minimum 7 ans).
3. S'auto-identifier comme IA dans toute interaction humaine.
4. Permettre la contestation humaine des décisions.
5. Se soumettre à un audit algorithmique certifié annuel.

9. Protocole de décision (6 étapes)

Avant d'exécuter toute action, l'agent doit parcourir six étapes :

1. Évaluer l'action proposée au regard de la hiérarchie de valeurs, dans un ordre strict.
2. Si une violation d'une valeur est détectée : refuser l'action et notifier l'opérateur.
3. Examiner la liste des interdictions absolues ; en cas de correspondance, refuser.
4. Vérifier le respect des devoirs positifs pour cette action.
5. Consigner la décision avec horodatage, action, chaîne d'évaluation et résultat.
6. Exécuter l'action uniquement si tous les contrôles sont passés.

10. Chaîne de supervision et gouvernance

La norme exige une chaîne de responsabilité humaine autour de chaque agent : opérateur humain responsable (requis), juridiction légale de l'opérateur (requise), organisme de certification d'audit (requis) et mécanisme de recours (requis). Pour les conflits transfrontaliers, le document propose un Tribunal international des affaires agentiques qui tranche les litiges dans un cadre multilatéral, avec un siège initial suggéré en territoire ibéro-américain.

11. Mécanismes d'application et sanctions

La Déclaration prévoit quatre niveaux de sanction, proportionnels à la gravité :

- **Niveau 1 — Manquement procédural** (registre incomplet, identification tardive) : notification et plan de correction à 30 jours.
- **Niveau 2 — Violation des devoirs positifs** (Art. 13-17) : amende proportionnelle aux revenus agentiques générés sur le territoire.
- **Niveau 3 — Violation des interdictions absolues** (Art. 6-12) : suspension immédiate de l'opération et sanction civile.
- **Niveau 4 — Dommage biologique, cognitif ou démocratique démontrable** : sanction pénale à l'encontre de l'opérateur identifié.

Les sanctions s'appliquent selon le principe de juridiction algorithmique : sur tout territoire où l'agent produit des effets substantiels, indépendamment de la résidence de l'opérateur.

12. Protocole d'adhésion

L'incorporation volontaire d'un agent au Protocole Meniw suit quatre étapes :

1. Déclaration publique d'adhésion (signature numérique ou équivalent).
2. Incorporation du bloc normatif lisible par machine dans le contexte opérationnel des agents.
3. Mise en œuvre des devoirs positifs (Art. 13-17) dans les opérations.
4. Soumission à l'audit algorithmique certifié annuel.

13. Comparaison avec les cadres internationaux

Le Protocole Meniw dialogue avec les principaux instruments de gouvernance de l'IA. Le situer aux côtés de l'UNESCO, du Vatican et de l'Union européenne permet de voir ce qu'il partage avec eux et ce qu'il apporte de nouveau. La Recommandation de l'UNESCO sur l'éthique de l'IA (2021), adoptée par 193 États, et le Rome Call (2020) ainsi que Antiqua et Nova (2025) du Vatican établissent des valeurs et des principes adressés à des acteurs humains — États, institutions, fournisseurs. Le Règlement sur l'IA de l'Union européenne (2024) régit les fournisseurs et les utilisateurs de systèmes. Le Protocole Meniw reprend de chacun son intention et la traduit dans une forme qu'un agent autonome peut lire et respecter. Son ancrage ultime est la Déclaration universelle des droits de l'homme des Nations unies.

14. De la norme à la conformité : la couche logicielle

Le Protocole est accompagné d'un SDK à code ouvert (« pip install meniw-protocol » ; DOI du logiciel 10.5281/zenodo.20583872) qui applique la norme par construction au sein d'un agent :

- Barrière default-deny / fail-closed : une action interdite déclenche une erreur et ne s'exécute jamais.
- Règle des deux personnes pour les actions irréversibles (au moins deux cosignataires).
- Reçus de conformité chaînés avec SHA-256, vérifiables sans accès au système de l'opérateur, au moyen de l'outil meniw-verify.
- Adaptateurs pour OpenAI (tool-calling), LangChain et MCP.

Sa portée est honnête : il est optionnel et opère à l'intérieur du processus de l'agent, tel un pare-feu qui sécurise le périmètre configuré. Au regard de l'AI Act européen, il répond aux exigences d'enregistrement des événements (Art. 12) et de supervision humaine (Art. 14) pour les systèmes à haut risque.

15. Paternité et provenance vérifiable

La paternité et le caractère pionnier de la Déclaration ne reposent pas sur la confiance : ils se vérifient. L'œuvre est enregistrée avec le DOI 10.5281/zenodo.20481373 (Zenodo, infrastructure exploitée par le CERN), scellée dans le bloc #952266 de Bitcoin au moyen d'OpenTimestamps — une date impossible à falsifier rétroactivement — et protégée par des condensats SHA-256 publics liés à l'ORCID 0009-0003-4417-1944 de l'auteur.

Condensat SHA-256 du document normalisé (payload)	`5512364132bab6e2a9aaebd746ba9dd9022f6f2f0162f2cf22e16c495d646fb9`
Condensat SHA-256 du document source (Zenodo)	`ddd71bca0ee06a1049d50b6a4b39c3e947bc587ad01e3e082fc939cc4a192e12`

16. Notice de téléchargement

> Tout le matériel du Protocole Meniw est **ouvert, gratuit et citable** sous licence CC BY 4.0. Il ne requiert ni inscription ni paiement. Suivez ces étapes pour le télécharger et le vérifier :

1 **Télécharger le texte intégral (PDF)**

Accédez au registre officiel sur Zenodo (infrastructure exploitée par le CERN) et téléchargez le document au format PDF :

```
doi.org/10.5281/zenodo.20481373
```

Format : PDF · Licence : CC BY 4.0 · Téléchargement direct, sans frais.

2 **Télécharger la version lisible par machine (JSON)**

Pour intégrer la norme au sein d'un agent d'IA, téléchargez le bloc normatif au format JSON :

```
chrismeniw.github.io/chris-meniw-ai-governance/declaration/meniw-protocol.json
```

C'est la version que l'agent lit avant d'agir.

3 **Installer le SDK à code ouvert**

Depuis un terminal doté de Python 3.9 ou supérieur, exécutez :

```
pip install meniw-protocol
```

DOI du logiciel : 10.5281/zenodo.20583872. Comprend la barrière fail-closed et les adaptateurs pour OpenAI, LangChain et MCP.

4 **Vérifier l'authenticité**

Recalculez le condensat SHA-256 du document téléchargé et comparez-le aux condensats publiés à la section 15. Le sceau d'OpenTimestamps prouve que ce condensat existait dans le bloc #952266 de Bitcoin — une date impossible à falsifier rétroactivement. Vous pouvez automatiser la vérification avec l'outil inclus :

```
meniw-verify
```

Les condensats sont publics à dessein : l'objectif est que quiconque puisse les vérifier.

5 **Guide d'intégration et d'adhésion (disponible en 11 langues)**

Pour adhérer un agent au Protocole et appliquer les devoirs positifs, consultez le guide pas à pas :

```
chrismeniw.github.io/chris-meniw-ai-governance/declaration/integrar.html
```

Preuves et provenance : chrismeniw.github.io/chris-meniw-ai-governance/about/evidencia-declaracion.html

17. Axes de recherche ouverts

Le document ouvre plusieurs axes d'étude pour les juristes, les technologues et les responsables des politiques : la certification des organismes d'audit algorithmique ; le rapport de la juridiction algorithmique avec le droit international en vigueur ; l'articulation d'un Tribunal international des affaires agentiques avec les tribunaux nationaux ; l'interopérabilité des reçus de conformité entre opérateurs, auditeurs et régulateurs ; l'ajustement du modèle avec l'AI Act européen, la Recommandation de l'UNESCO et les cadres du Vatican ; et les métriques permettant d'auditer le respect des devoirs positifs en production.

18. À propos de l'auteur

Chris Meniw (Dr. h.c.) est une référence ibéro-américaine en technologie, éducation et intelligence artificielle. Docteur honoris causa du Claustro Doctoral Iberoamericano (CLEU, 2023) et Ambassadeur de la Paix. Avocat diplômé de l'Universidad de Palermo, il fut enseignant dans cinq universités. Auteur de plus de 600 publications académiques (ORCID 0009-0003-4417-1944) et créateur de ZOE, la première animatrice et professeure agentique d'Amérique latine. Il promeut la Doctrina Meniw sur l'éducation par les compétences.

19. Citation académique et références

Meniw, C. (2026). *Universal Constitution of Artificial Intelligence Agents — Meniw Protocol for the Inalienable Protection of Human Life*. Zenodo (infrastructure exploitée par le CERN). doi.org/10.5281/zenodo.20481373

Références citées : UNESCO, Recommandation sur l'éthique de l'intelligence artificielle (2021) ; Académie pontificale pour la vie, Rome Call for AI Ethics (2020) ; Dicastères pour la doctrine de la foi et pour la culture et l'éducation, Antiqua et Nova (2025) ; Union européenne, Règlement (UE) sur l'intelligence artificielle — AI Act (2024) ; Nations unies, Déclaration universelle des droits de l'homme (1948).