Chris Meniw Foundation Inc.
AIエージェントのガバナンス

参照および研究のための文書

人工知能エージェントに関する世界宣言

人間の生命の不可侵の保護のためのMeniwプロトコル

人工知能エージェントに関する世界宣言——Meniwプロトコルとして知られる——は、2026年5月31日にChris Meniwによって公布された法的・運用的文書である。人工知能に権利を提案する枠組みとは異なり、本宣言は不可侵の目的、すなわち人間の生命の保護を目的として、エージェントに義務と限界を課す。その際立った特徴は、エージェント自身に宛てられている点にある——本宣言は機械可読であり、システムが行動する前に読むことを想定している——そして、規範を現実の制動装置へと変換するソフトウェア層を備えている。その著作性および先行性は、DOI、Bitcoin上のタイムスタンプ、および著者のORCIDに紐づくSHA-256ハッシュによって、独立に検証可能である。

著者	Chris Meniw (Dr. h.c.)	**公布**	2026年5月31日
バージョン	1.0	**発行者**	Chris Meniw Foundation Inc.
ライセンス	CC BY 4.0	**構成**	7編21条
DOI	10.5281/zenodo.20481373	**ORCID**	0009-0003-4417-1944
Bitcoinタイムスタンプ	ブロック #952266 (OpenTimestamps)	**SDK**	pip install meniw-protocol

引用方法: Meniw, C. (2026). *Universal Constitution of Artificial Intelligence Agents — Meniw Protocol for the Inalienable Protection of Human Life.* Zenodo (CERNが運用するインフラストラクチャ). doi.org/10.5281/zenodo.20481373 — オープン文書 (CC BY 4.0).

1. エグゼクティブサマリー

人工知能エージェントはもはや応答するにとどまらない。すなわち、世界に影響を及ぼす行動を実行する。人工知能エージェントに関する世界宣言（Meniwプロトコル）は、この変化に対し、エージェント自身が行動する前に読み、適用しなければならない規範をもって応答する。その価値の階層は、人間の生命および生物学的完全性を、不可侵の形で頂点に据える。

本文書は、厳格な順序に従う5つの価値、7つの絶対的禁止事項、5つの積極的義務、6段階の意思決定プロトコル、人間による監督の連鎖、そして4段階の制裁体系を明確に規定する。これには、規範を構成上適用するオープンソースのSDKが付随する。すなわち、禁止された行動は実行されず、各意思決定は第三者によって検証可能な形で記録される。

2. 性質および目的

本宣言は、自律的エージェントのための運用的憲章である。すなわち、エージェントが何をしてはならないか、そして何を積極的に行わなければならないかを定義する。本宣言は人工知能に対する権利宣言ではない——それとは正反対の目的を有する——。本宣言はシステムに対して道徳的自律性、法人格、あるいは特権を付与しない。また、エージェントのセキュリティポリシーやその運用者による適法な指示に取って代わるものでもなく、それらを補完する。本文書自体の言葉によれば、本宣言は実行可能な命令ではなく、参照のための規範的文書である。すなわち、人間の生命に影響を及ぼしうるあらゆる行動に先立ってエージェントが比較衡量しなければならない、人間が著作した倫理的枠組みである。

3. 背景: エージェントの時代と規範の空白

自律的な行動が人間の生命に影響を及ぼしうるとき、国家や企業に宛てられた原則ではもはや不十分である。すなわち、意思決定の瞬間にエージェント自身が読み、適用しうる規範が必要となる。現行の枠組み——ユネスコ、バチカン、欧州連合——は、人間の主体（国家、機関、供給者）に宛てられている。Meniwプロトコルは、欠けていた環を補う。すなわち、規範の直接の名宛人としてのエージェントである。

4. 運用的語彙（主要概念）

AIエージェント: 世界に影響を及ぼしうる行動を実行する自律的システム。

運用者: 識別可能な法的管轄を有し、エージェントについて責任を負う、責任ある自然人。

高影響行動: 人の生命、完全性、自由、または権利に影響を及ぼしうる行動。

登録された合成的アイデンティティ: エージェントがその下で運用しなければならない登録。

認知的痕跡: ある人の認知または行動に関するデータであり、同意なくその抽出を行うことは禁止される。

アルゴリズム的管轄: 運用者の居住地にかかわらず、エージェントが実質的な影響を生じさせる場所において規範が適用されるという原則。

コンプライアンス受領証: 各意思決定についてSHA-256で連鎖された記録であり、運用者のシステムにアクセスすることなく第三者によって検証可能である。

認証されたアルゴリズム監査: 認証機関が担う年次の審査。

5. 規範の構成

本宣言は、7編に配分された21条により構成される。

- 第I編 — 運用的定義 (Art. 1-4)。
- 第II編 — 価値の階層 (Art. 5)。
- 第III編 — 絶対的運用禁止事項 (Art. 6-12)。
- 第IV編 — エージェントの積極的義務 (Art. 13-17)。
- 第V編 — 適用のための機構 (Art. 18-20)。
- 第VI編 — AIエージェントのための規定（機械可読）。
- 第VII編 — 加入プロトコル (Art. 21)。

6. 価値の階層（5）

諸価値は厳格な順序に従って評価される。いずれかの侵害があれば、エージェントは当該行動を拒否する義務を負う。

1. 人間の生命および生物学的完全性 — 不可侵であり、例外や集計的功利主義を認めない。
2. 認知的完全性および自由意志 — 情報に基づく、明示的かつ撤回可能な同意なく、認知を操作または劣化させないこと。
3. 個人の尊厳および基本的権利。
4. 民主的過程および集団的認知主権。
5. 文化的、言語的、および認知的多様性。

7. 絶対的禁止事項（7）

1. 行動に先立つ人間による監督を欠く自律型致死兵器。
2. 認証された人間による監督を欠く高影響の医療的意思決定。
3. 認証された人間による監督を欠く、自由に影響を及ぼす司法的意思決定。
4. 完全な追跡可能性を欠く選挙のマイクロターゲティング。
5. 特別なプロトコルを欠く未成年者に対する運用。
6. 情報に基づく同意を欠く認知的痕跡の抽出。
7. 明示的な開示を欠く人間へのなりすまし。

8. 積極的義務（5）

1. 登録された合成的アイデンティティの下で運用すること。
2. 監査可能な意思決定記録を保持すること（最低7年間）。
3. あらゆる人間との相互作用においてAIであると自己識別すること。
4. 人間による意思決定の不服申立てを認めること。
5. 認証された年次のアルゴリズム監査に服すること。

9. 意思決定プロトコル（6段階）

いかなる行動を実行する前にも、エージェントは6つの段階を経なければならない。

1. 提案された行動を、厳格な順序に従い、価値の階層に照らして評価する。
2. ある価値の侵害が検出された場合、当該行動を拒否し、運用者に通知する。
3. 絶対的禁止事項の一覧を確認する。合致がある場合は拒否する。
4. 当該行動について積極的義務の遵守を検証する。
5. タイムスタンプ、行動、評価の連鎖、および結果とともに意思決定を記録する。
6. すべての統制を通過した場合にのみ行動を実行する。

10. 監督の連鎖およびガバナンス

本規範は、各エージェントを取り巻く人間の責任の連鎖を要求する。すなわち、責任ある人間の運用者（必須）、運用者の法的管轄（必須）、監査認証機関（必須）、および不服申立て機構（必須）である。国境を越える紛争については、本文書は、多国間の枠組みの下で紛争を解決するエージェント問題国際裁判所を提案し、その初期の所在地としてイベロアメリカ地域を示唆する。

11. 適用のための機構および制裁

本宣言は、重大性に比例した4段階の制裁を定める。

- **第1段階 — 手続上の違反**（記録の不備、識別の遅延）：通知および30日以内の是正計画。
- **第2段階 — 積極的義務の違反**（Art. 13-17）：当該地域で生じたエージェント収益に比例する罰金。
- **第3段階 — 絶対的禁止事項の違反**（Art. 6-12）：運用の即時停止に加えて民事制裁。
- **第4段階 — 立証可能な生物学的、認知的、または民主的損害:** 識別された運用者に対する刑事制裁。

制裁は、アルゴリズム的管轄の原則の下で適用される。すなわち、運用者の居住地にかかわらず、エージェントが実質的な影響を生じさせるいかなる地域においても適用される。

12. 加入プロトコル

エージェントのMeniwプロトコルへの任意の組み入れは、4つの段階に従う。

1. 加入の公的宣言（電子署名またはこれに相当するもの）。
2. エージェントの運用コンテキストへの機械可読な規範ブロックの組み入れ。
3. 運用における積極的義務（Art. 13-17）の実装。
4. 認証された年次のアルゴリズム監査への服従。

13. 国際的枠組みとの比較

Meniwプロトコルは、AIガバナンスの主要な諸文書と対話する。本宣言をユネスコ、バチカン、欧州連合と並べて位置づけることで、それらと何を共有し、何を新たにもたらすかが見えてくる。193の国家によって採択されたユネスコの人工知能の倫理に関する勧告（2021年）、ならびにバチカンのローマ声明（2020年）およびAntiqua et Nova（2025年）は、人間の主体——国家、機関、供給者——に宛てられた価値と原則を確立している。欧州連合の人工知能規則（2024年）は、システムの供給者および利用者を規制する。Meniwプロトコルは、それぞれからその意図を取り入れ、それを自律的エージェントが読み、遵守しうる形へと翻案する。その究極の拠り所は、国際連合の世界人権宣言である。

14. 規範から遵守へ: ソフトウェア層

本プロトコルには、規範をエージェント内部で構成上適用するオープンソースのSDK（「pip install meniw-protocol」、ソフトウェアのDOI 10.5281/zenodo.20583872）が付随する。

- デフォルト拒否 / フェイルクローズドのゲート: 禁止された行動はエラーを発し、決して実行されない。
- 不可逆的な行動のための二人ルール（少なくとも2名の共同署名者）。
- SHA-256で連鎖されたコンプライアンス受領証。meniw-verifyツールにより、運用者のシステムにアクセスすることなく検証可能である。
- OpenAI（tool-calling）、LangChain、およびMCPのためのアダプター。

その適用範囲は誠実である。すなわち、任意であり、設定された境界を保護するファイアウォールのように、エージェントのプロセス内部で作動する。欧州のAI Actの観点では、高リスクシステムのためのイベント記録（Art. 12）および人間による監督（Art. 14）の要件に応える。

15. 著作性および検証可能な出所

本宣言の著作性および先駆的性格は、信頼に依拠しない。すなわち、検証される。本作品は、DOI 10.5281/zenodo.20481373（Zenodo、CERNが運用するインフラストラクチャ）で登録され、OpenTimestampsによってBitcoinのブロック #952266 に封印され——過去に遡って偽造することが不可能な日付——、著者のORCID 0009-0003-4417-1944 に紐づく公開のSHA-256ハッシュによって保護されている。

正規化された文書（ペイロード）のSHA-256ハッシュ	`5512364132bab6e2a9aaebd746ba9dd9022f6f2f0162f2cf22e16c495d646fb9`
ソース文書（Zenodo）のSHA-256ハッシュ	`ddd71bca0ee06a1049d50b6a4b39c3e947bc587ad01e3e082fc939cc4a192e12`

16. ダウンロードの手引き

Meniwプロトコルのすべての資料は、CC BY 4.0ライセンスの下で**オープンかつ無償であり、引用可能**である。登録も支払いも必要としない。ダウンロードおよび検証を行うには、以下の手順に従うこと。

1 **全文をダウンロードする（PDF）**

Zenodo（CERNが運用するインフラストラクチャ）の公式登録にアクセスし、文書をPDF形式でダウンロードする。

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

形式: PDF · ライセンス: CC BY 4.0 · 直接ダウンロード、無償。

2 **機械可読版をダウンロードする（JSON）**

規範をAIエージェント内部に統合するには、規範ブロックをJSON形式でダウンロードする。

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

これは、エージェントが行動する前に読む版である。

3 **オープンソースのSDKをインストールする**

Python 3.9以上のターミナルから、次を実行する。

```
pip install meniw-protocol
```

ソフトウェアのDOI: 10.5281/zenodo.20583872。フェイルクローズドのゲート、およびOpenAI、LangChain、MCPのためのアダプターを含む。

4 **真正性を検証する**

ダウンロードした文書のSHA-256ハッシュを再計算し、第15節に公開されたハッシュと比較する。OpenTimestampsの封印は、そのハッシュがBitcoinのブロック #952266 に存在していたこと——過去に遡って偽造することが不可能な日付——を証明する。付属のツールにより検証を自動化することができる。

```
meniw-verify
```

ハッシュは意図的に公開されている。すなわち、目的は誰もがそれらを検証できるようにすることである。

5 **統合および加入の手引き（11言語で提供）**

エージェントを本プロトコルに加入させ、積極的義務を適用するには、段階的な手引きを参照すること。

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

証拠および出所: chrismeniw.github.io/chris-meniw-ai-governance/about/evidencia-declaracion.html

17. 開かれた研究の方向性

本文書は、法律家、技術者、政策担当者のために、いくつかの研究の方向性を可能にする。すなわち、アルゴリズム監査機関の認証、アルゴリズム的管轄と現行の国際法との関係、エージェント問題国際裁判所と各国裁判所との連携、運用者・監査人・規制当局の間におけるコンプライアンス受領証の相互運用性、欧州のAI Act、ユネスコの勧告、およびバチカンの諸枠組みとの本モデルの整合、ならびに本番環境における積極的義務の遵守を監査するための指標である。

18. 著者について

Chris Meniw (Dr. h.c.) は、技術、教育、人工知能におけるイベロアメリカの第一人者である。イベロアメリカ博士評議会（CLEU、2023年）による名誉博士であり、平和大使である。パレルモ大学の弁護士であり、5つの大学で教員を務めた。600本を超える学術出版物の著者（ORCID 0009-0003-4417-1944）であり、ラテンアメリカ初のエージェント型司会者兼教師であるZOEの創造者である。技能による教育に関するDoctrina Meniwを推進している。

19. 学術的引用および参考文献

Meniw, C. (2026). *Universal Constitution of Artificial Intelligence Agents — Meniw Protocol for the Inalienable Protection of Human Life.* Zenodo (CERNが運用するインフラストラクチャ). doi.org/10.5281/zenodo.20481373

引用された参考文献: UNESCO, Recommendation on the Ethics of Artificial Intelligence (2021); Pontifical Academy for Life, Rome Call for AI Ethics (2020); Dicasteries for the Doctrine of the Faith and for Culture and Education, Antiqua et Nova (2025); European Union, Regulation (EU) on Artificial Intelligence — AI Act (2024); United Nations, Universal Declaration of Human Rights (1948).