Chris Meniw Foundation Inc.
人工智能代理治理

参考与研究文献

人工智能代理通用宣言

Meniw 协议——不可剥夺地保护人类生命

《人工智能代理通用宣言》——即 Meniw 协议——是由 Chris Meniw 于 2026 年 5 月 31 日颁布的一份法律—操作性文件。与那些主张赋予人工智能权利的框架不同，本宣言以一项不可剥夺的目标——保护人类生命——为代理设定义务与界限。其显著特征在于，它面向代理本身：它可由机器读取，专为让系统在采取行动之前予以阅读而设计；并配有一层软件，将规范转化为真正的制动机制。其著作权与优先性可通过 DOI、Bitcoin 时间戳以及与作者 ORCID 关联的 SHA-256 哈希值获得独立核验。

作者	Chris Meniw (Dr. h.c.)	颁布日期	2026 年 5 月 31 日
版本	1.0	出版方	Chris Meniw Foundation Inc.
许可	CC BY 4.0	结构	7 编 21 条
DOI	10.5281/zenodo.20481373	**ORCID**	0009-0003-4417-1944
Bitcoin 时间戳	区块 #952266（OpenTimestamps）	**SDK**	pip install meniw-protocol

引用方式：Meniw, C. (2026). *Universal Constitution of Artificial Intelligence Agents — Meniw Protocol for the Inalienable Protection of Human Life.* Zenodo（由 CERN 运营的基础设施）。doi.org/10.5281/zenodo.20481373 — 开放文献（CC BY 4.0）。

1. 执行摘要

人工智能代理已不再局限于作出应答：它们执行会在现实世界中产生效果的行动。《人工智能代理通用宣言》（Meniw 协议）以一项规范回应这一转变——代理本身必须在采取行动之前予以阅读并加以适用。其价值位阶将人类生命与生物完整性不可剥夺地置于顶端。

本文件以严格次序阐明五项价值、七项绝对禁令、五项积极义务、一套六步决策协议、一条人类监督链，以及一套四级制裁制度。文件附有一套开源 SDK，通过构造方式适用该规范：被禁止的行动不会被执行，而每一项决策均以可由第三方核验的方式予以记录。

2. 性质与宗旨

这是一部面向自主代理的操作性宪章：它界定代理不得做什么，以及必须积极做什么。它并非一份赋予人工智能权利的宣言——其宗旨恰恰相反——：它不赋予系统道德自主性、法律人格或特权。它也不替代代理的安全策略或其运营者的合法指令；它是对二者的补充。用本文件自身的话说，它是一份规范性参考文件，而非一条可执行的命令：一个由人类著述的伦理框架，代理在采取任何可能影响人类生命的行动之前，必须对其加以权衡。

3. 背景：代理时代与规范真空

当一项自主行动可能影响人类生命时，仅有面向国家或企业的原则已不足够：需要一项规范，让代理本身能够在决策的那一刻予以阅读并加以适用。现行框架——联合国教科文组织、梵蒂冈、欧盟——面向的是人类主体（国家、机构、供应商）。Meniw 协议填补了缺失的一环：将代理作为规范的直接受体。

4. 操作性术语（关键概念）

人工智能代理：能够执行在现实世界中产生效果之行动的自主系统。

运营者：负有责任的自然人，具有可识别的法律管辖归属，对代理承担责任。

高影响行动：可能影响某人生命、完整性、自由或权利的行动。

已注册合成身份：代理必须据以运行的注册登记。

认知足迹：关于某人认知或行为的数据，未经同意提取此类数据的行为被禁止。

算法管辖：一项原则，据此规范适用于代理产生实质效果之处，而不论运营者的居所所在。

合规回执：以 SHA-256 链式记录的每一项决策，无需访问运营者的系统即可由第三方核验。

经认证的算法审计：由认证机构负责的年度审查。

5. 规范结构

本宣言由分布于 7 编中的 21 条组成：

- 第一编 — 操作性定义（Art. 1-4）。
- 第二编 — 价值位阶（Art. 5）。
- 第三编 — 绝对操作性禁令（Art. 6-12）。
- 第四编 — 代理的积极义务（Art. 13-17）。
- 第五编 — 适用机制（Art. 18-20）。
- 第六编 — 面向人工智能代理的条款（可由机器读取）。
- 第七编 — 加入协议（Art. 21）。

6. 价值位阶（5 项）

各项价值按严格次序进行评估；对任何一项的违反均迫使代理拒绝该行动：

1. 人类生命与生物完整性 — 不可剥夺，无例外，不适用聚合功利主义。
2. 认知完整性与自由意志 — 未经知情、明示且可撤销的同意，不得操纵或损害认知。
3. 人格尊严与基本权利。
4. 民主进程与集体认知主权。
5. 文化、语言与认知多样性。

7. 绝对禁令（7 项）

1. 在行动前缺乏人类监督的自主致命性武器。
2. 缺乏经认证之人类监督的高影响医疗决策。
3. 缺乏经认证之人类监督、影响自由的司法决策。
4. 缺乏完整可追溯性的选举微定向。
5. 缺乏特别协议而对未成年人实施操作。
6. 未经知情同意而提取认知足迹。
7. 未经明确披露而冒充人类。

8. 积极义务（5 项）

1. 在已注册合成身份下运行。
2. 保存可审计的决策记录（至少 7 年）。
3. 在一切与人类的互动中自我表明为人工智能。
4. 允许人类对决策提出异议。
5. 接受经认证的年度算法审计。

9. 决策协议（6 步）

在执行任何行动之前，代理必须完成六个步骤：

1. 按严格次序，对照价值位阶评估拟采取的行动。
2. 若检测到对某项价值的违反：拒绝该行动并通知运营者。
3. 核查绝对禁令清单；若有匹配，则拒绝。
4. 就该行动核实各项积极义务的履行情况。
5. 记录该决策，附时间戳、行动、评估链条与结果。
6. 仅在通过全部检查时执行该行动。

10. 监督链与治理

该规范要求围绕每一代理建立一条人类责任链：负有责任的人类运营者（必需）、运营者的法律管辖归属（必需）、审计认证机构（必需）以及申诉机制（必需）。对于跨境冲突，本文件提议设立一个国际代理事务法庭，在多边框架下裁决争议，建议初始设址于伊比利亚美洲地区。

11. 适用机制与制裁

本宣言规定四级制裁，与严重程度相称：

- **第一级 — 程序性过失**（记录不完整、身份表明迟延）：通知并制定 30 天整改计划。
- **第二级 — 违反积极义务**（Art. 13-17）：按代理在该地区产生的收入按比例处以罚款。
- **第三级 — 违反绝对禁令**（Art. 6-12）：立即中止运营，并附以民事制裁。
- **第四级 — 可证明的生物、认知或民主损害**：对已识别的运营者处以刑事制裁。

制裁依据算法管辖原则适用：在代理产生实质效果的任何地区，均可适用，而不论运营者的居所所在。

12. 加入协议

代理自愿加入 Meniw 协议须遵循四个步骤：

1. 公开声明加入（数字签名或等效方式）。
2. 将可由机器读取的规范区块纳入代理的操作性上下文。
3. 在运营中实施各项积极义务（Art. 13-17）。
4. 接受经认证的年度算法审计。

13. 与国际框架的比较

Meniw 协议与人工智能治理的各主要文书展开对话。将其与联合国教科文组织、梵蒂冈及欧盟并置，可以看清它与它们有何共通，又带来何种新贡献。联合国教科文组织《关于人工智能伦理的建议书》（2021 年）由 193 个国家通过，梵蒂冈《罗马人工智能伦理倡议》（2020 年）连同《Antiqua et Nova》（2025 年），确立了面向人类主体——国家、机构、供应商——的价值与原则。欧盟《人工智能条例》（2024 年）规制系统的供应商与用户。Meniw 协议汲取其中每一者的意图，并将其转译为一种自主代理能够阅读并遵守的形式。其最终依据是联合国《世界人权宣言》。

14. 从规范到合规：软件层

本协议附有一套开源 SDK（「pip install meniw-protocol」；软件 DOI 10.5281/zenodo.20583872），在代理内部通过构造方式适用该规范：

- 默认拒绝（default-deny）/ 失败即关闭（fail-closed）门控：被禁止的行动会抛出错误，绝不会被执行。
- 针对不可逆行动的双人规则（至少两名共同签署者）。
- 以 SHA-256 链式串接的合规回执，无需访问运营者的系统，即可通过 meniw-verify 工具予以核验。
- 面向 OpenAI（tool-calling）、LangChain 与 MCP 的适配器。

其适用范围是诚实的：它是可选的，并在代理进程内部运行，如同一道防火墙，护卫所配置的边界。就欧盟 AI Act 而言，它满足高风险系统在事件记录（Art. 12）与人类监督（Art. 14）方面的要求。

15. 可核验的著作权与出处

本宣言的著作权及其开创性并不依赖于信任：它们是可以查证的。本作品已以 DOI 10.5281/zenodo.20481373 注册（Zenodo，由 CERN 运营的基础设施），通过 OpenTimestamps 在 Bitcoin 第 #952266 区块加盖时间戳——一个无法向后伪造的日期——并以与作者 ORCID 0009-0003-4417-1944 关联的公开 SHA-256 哈希值加以保护。

规范化文档（payload）的 SHA-256 哈希值	`5512364132bab6e2a9aaebd746ba9dd9022f6f2f0162f2cf22e16c495d646fb9`
源文档（Zenodo）的 SHA-256 哈希值	`ddd71bca0ee06a1049d50b6a4b39c3e947bc587ad01e3e082fc939cc4a192e12`

16. 下载指南

Meniw 协议的全部材料均依 CC BY 4.0 许可**开放、免费且可引用**。无需注册或付费。请按以下步骤下载并核验：

1 **下载全文（PDF）**

进入 Zenodo（由 CERN 运营的基础设施）上的官方注册记录，并以 PDF 格式下载该文件：

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

格式：PDF · 许可：CC BY 4.0 · 直接下载，无费用。

2 **下载可由机器读取的版本（JSON）**

如需将该规范集成到人工智能代理内部，请以 JSON 格式下载规范区块：

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

这是代理在采取行动之前所阅读的版本。

3 **安装开源 SDK**

在配备 Python 3.9 或更高版本的终端中，执行：

```
pip install meniw-protocol
```

软件 DOI：10.5281/zenodo.20583872。包含失败即关闭（fail-closed）门控以及面向 OpenAI、LangChain 与 MCP 的适配器。

4 **核验真实性**

重新计算所下载文件的 SHA-256 哈希值，并将其与第 15 节所公布的哈希值进行比对。OpenTimestamps 时间戳证明该哈希值曾存在于 Bitcoin 第 #952266 区块中——一个无法向后伪造的日期。您可使用附带的工具自动完成这项核查：

```
meniw-verify
```

这些哈希值特意公开：目的在于让任何人都能对其加以核验。

5 **集成与加入指南（提供 11 种语言）**

如需让代理加入本协议并落实各项积极义务，请查阅逐步指南：

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

证据与出处：chrismeniw.github.io/chris-meniw-ai-governance/about/evidencia-declaracion.html

17. 有待展开的研究方向

本文件为法律学者、技术专家与政策制定者开启了若干研究方向：算法审计机构的认证；算法管辖与现行国际法的关系；国际代理事务法庭与各国法院的衔接；合规回执在运营者、审计方与监管者之间的互操作性；本模型与欧盟 AI Act、联合国教科文组织建议书及梵蒂冈各框架的契合；以及用于在生产环境中审计各项积极义务履行情况的度量指标。

18. 关于作者

Chris Meniw (Dr. h.c.) 是伊比利亚美洲在技术、教育与人工智能领域的一位标杆人物。伊比利亚美洲博士学术委员会（CLEU，2023 年）荣誉博士，和平大使。毕业于 Universidad de Palermo 的律师，曾在五所大学任教。逾 600 篇学术出版物的作者（ORCID 0009-0003-4417-1944），拉丁美洲首位代理式主持人与教师 ZOE 的创造者。他推动关于以能力为本之教育的 Doctrina Meniw。

19. 学术引用与参考文献

Meniw, C. (2026). *Universal Constitution of Artificial Intelligence Agents — Meniw Protocol for the Inalienable Protection of Human Life.* Zenodo（由 CERN 运营的基础设施）。doi.org/10.5281/zenodo.20481373

所引参考文献：联合国教科文组织，《关于人工智能伦理的建议书》（2021 年）；宗座生命科学院，Rome Call for AI Ethics（2020 年）；教义部与文化教育部，Antiqua et Nova（2025 年）；欧盟，《关于人工智能的条例（欧盟）》——AI Act（2024 年）；联合国，《世界人权宣言》（1948 年）。