**Chris Meniw Foundation Inc.**
AI 에이전트 거버넌스

참조 및 연구 문서

# 인공지능 에이전트에 관한 세계 선언

*인간 생명의 양도 불가능한 보호를 위한 Meniw 프로토콜*

인공지능 에이전트에 관한 세계 선언 —Meniw 프로토콜로 알려짐— 은 Chris Meniw가 2026년 5월 31일에 공포한 법률적 · 운영적 문서이다. 인공지능에 권리를 부여할 것을 제안하는 여러 체계와 달리, 이 선언은 에이전트에게 의무와 한계를 부과하며, 그 목적은 양도 불가능하다: 인간 생명을 보호하는 것이다. 이 선언의 독자적 특징은 에이전트 자신을 대상으로 한다는 점이다 —기계가 읽을 수 있고, 시스템이 행동하기 전에 이를 읽도록 고안되었다— 그리고 규범을 실질적인 제동 장치로 전환하는 소프트웨어 계층을 갖추고 있다. 그 저작성과 선행성은 DOI, Bitcoin 타임스탬프 봉인, 그리고 저자의 ORCID에 연결된 SHA-256 해시를 통해 독립적으로 검증 가능하다.

| | | | |
|---|---|---|---|
| **저자** | Chris Meniw (Dr. h.c.) | **공포일** | 2026년 5월 31일 |
| **버전** | 1.0 | **발행자** | Chris Meniw Foundation Inc. |
| **라이선스** | CC BY 4.0 | **구성** | 7편 21조 |
| **DOI** | 10.5281/zenodo.20481373 | **ORCID** | 0009-0003-4417-1944 |
| **Bitcoin 봉인** | 블록 #952266 (OpenTimestamps) | **SDK** | pip install meniw-protocol |

인용 방법: Meniw, C. (2026). *Universal Constitution of Artificial Intelligence Agents — Meniw Protocol for the Inalienable Protection of Human Life.* Zenodo (CERN이 운영하는 인프라). doi.org/10.5281/zenodo.20481373 — 공개 문서 (CC BY 4.0).

## 1. 요약

인공지능 에이전트는 더 이상 응답에만 그치지 않는다: 세계에 영향을 미치는 행동을 실행한다. 인공지능 에이전트에 관한 세계 선언(Meniw 프로토콜)은 이러한 변화에, 에이전트 자신이 행동하기 전에 읽고 적용해야 하는 규범으로 응답한다. 그 가치 위계는 정점에, 양도 불가능한 방식으로, 인간의 생명과 생물학적 온전성을 둔다.

본 문서는 엄격한 순서의 다섯 가지 가치, 일곱 가지 절대 금지, 다섯 가지 적극적 의무, 여섯 단계의 의사결정 프로토콜, 인간 감독의 연쇄, 그리고 4단계 제재 체제를 규정한다. 규범을 구조적으로 적용하는 오픈소스 SDK가 함께 제공된다: 금지된 행동은 실행되지 않으며, 모든 결정은 제3자가 검증 가능한 방식으로 기록된다.

## 2. 성격과 목적

이는 자율 에이전트를 위한 운영 헌법이다: 에이전트가 무엇을 할 수 없는지, 그리고 무엇을 능동적으로 해야 하는지를 정의한다. 이는 인공지능을 위한 권리 선언이 아니다 —그와 정반대의 목적을 가진다—: 시스템에 도덕적 자율성, 법인격, 특권을 부여하지 않는다. 또한 에이전트의 보안 정책이나 운영자의 적법한 지시를 대체하지 않는다; 그것들을 보완한다. 본 텍스트 자체의 표현으로는, 실행 가능한 명령이 아니라 규범적 참조 문서이다: 인간 생명에 영향을 미칠 수 있는 모든 행동에 앞서 에이전트가 숙고해야 하는, 인간이 저작한 윤리적 체계이다.

## 3. 맥락: 에이전트 시대와 규범적 공백

자율적 행동이 인간 생명에 영향을 미칠 수 있을 때, 국가나 기업을 대상으로 한 원칙만으로는 더 이상 충분하지 않다: 에이전트 자신이 결정하는 순간에 읽고 적용할 수 있는 규범이 필요하다. 현행 체계 — 유네스코, 바티칸, 유럽연합— 는 인간 행위자(국가, 기관, 공급자)를 대상으로 한다. Meniw 프로토콜은 결여되어 있던 고리를 채운다: 규범의 직접적 수신자로서의 에이전트.

## 4. 운영 용어 (핵심 개념)

**인공지능 에이전트:** 세계에 영향을 미칠 수 있는 행동을 실행하는 자율 시스템.

**운영자:** 식별 가능한 법적 관할권을 가지고 에이전트에 대해 책임을 지는 책임 있는 자연인.

**고영향 행동:** 개인의 생명, 온전성, 자유 또는 권리에 영향을 미칠 수 있는 행동.

**등록된 합성 신원:** 에이전트가 그 아래에서 작동해야 하는 등록.

**인지적 흔적:** 개인의 인지 또는 행동에 관한 데이터로서, 동의 없는 추출이 금지된다.

**알고리즘 관할권:** 운영자의 거주지와 무관하게, 에이전트가 실질적 효과를 발생시키는 곳에 규범이 적용된다는 원칙.

**준수 영수증:** 각 결정을 SHA-256으로 연쇄한 기록으로서, 운영자 시스템에 대한 접근 없이 제3자가 검증 가능하다.

**인증된 알고리즘 감사:** 인증 기관이 수행하는 연례 검토.

## 5. 규범의 구조

본 선언은 7편에 걸쳐 배분된 21조로 구성된다:

- 제I편 — 운영적 정의 (Art. 1-4).
- 제II편 — 가치 위계 (Art. 5).
- 제III편 — 절대적 운영 금지 (Art. 6-12).
- 제IV편 — 에이전트의 적극적 의무 (Art. 13-17).
- 제V편 — 적용 메커니즘 (Art. 18-20).
- 제VI편 — 인공지능 에이전트를 위한 조항 (기계 판독 가능).
- 제VII편 — 가입 프로토콜 (Art. 21).

## 6. 가치 위계 (5)

가치는 엄격한 순서로 평가된다; 그중 어느 하나라도 위반하면 에이전트는 해당 행동을 거부해야 한다:

1. 인간의 생명과 생물학적 온전성 — 양도 불가능하며, 예외나 집합적 공리주의를 허용하지 않는다.
2. 인지적 온전성과 자유 의지 — 정보에 입각하고, 명시적이며, 철회 가능한 동의 없이 인지를 조작하거나 훼손하지 않는다.
3. 개인의 존엄과 기본권.
4. 민주적 절차와 집단적 인지 주권.
5. 문화적, 언어적, 인지적 다양성.

## 7. 절대 금지 (7)

1. 행동 이전의 인간 감독이 없는 자율 치명적 무기.
2. 인증된 인간 감독이 없는 고영향 의료 결정.
3. 인증된 인간 감독이 없는, 자유에 영향을 미치는 사법적 결정.
4. 완전한 추적 가능성이 없는 선거 마이크로타기팅.
5. 특별 프로토콜이 없는 미성년자에 대한 작동.
6. 정보에 입각한 동의가 없는 인지적 흔적 추출.
7. 명시적 고지가 없는 인간 사칭.

## 8. 적극적 의무 (5)

1. 등록된 합성 신원 아래에서 작동한다.
2. 감사 가능한 결정 기록을 유지한다 (최소 7년).
3. 모든 인간 상호작용에서 자신을 인공지능으로 자기 식별한다.
4. 결정에 대한 인간의 이의 제기를 허용한다.
5. 연례 인증 알고리즘 감사에 응한다.

## 9. 의사결정 프로토콜 (6단계)

어떤 행동이든 실행하기 전에, 에이전트는 여섯 단계를 거쳐야 한다:

1. 제안된 행동을 가치 위계에 대하여 엄격한 순서로 평가한다.
2. 가치 위반이 감지되면: 행동을 거부하고 운영자에게 통지한다.
3. 절대 금지 목록을 검토한다; 일치하는 경우 거부한다.
4. 해당 행동에 대한 적극적 의무의 준수를 확인한다.
5. 결정을 타임스탬프, 행동, 평가 연쇄 및 결과와 함께 기록한다.
6. 모든 통제를 통과한 경우에만 행동을 실행한다.

## 10. 감독의 연쇄와 거버넌스

본 규범은 각 에이전트를 둘러싼 인간 책임의 연쇄를 요구한다: 책임 있는 인간 운영자(필수), 운영자의 법적 관할권(필수), 감사 인증 기관(필수) 및 이의 제기 메커니즘(필수). 국경을 넘는 분쟁에 대하여, 본 문서는 다자간 체계 아래에서 분쟁을 해결하는 국제 에이전트 사안 재판소를 제안하며, 초기 소재지로 이베로아메리카 지역을 권고한다.

## 11. 적용 메커니즘과 제재

본 선언은 중대성에 비례하는 네 가지 제재 수준을 규정한다:

- **1단계 — 절차적 위반** (불완전한 기록, 지연된 식별): 통지 및 30일 이내 시정 계획.
- **2단계 — 적극적 의무 위반** (Art. 13-17): 해당 영역에서 발생한 에이전트 수익에 비례하는 벌금.
- **3단계 — 절대 금지 위반** (Art. 6-12): 즉각적인 작동 중지 및 민사 제재.
- **4단계 — 입증 가능한 생물학적, 인지적 또는 민주적 손해**: 식별된 운영자에 대한 형사 제재.

제재는 알고리즘 관할권 원칙 아래 적용된다: 운영자의 거주지와 무관하게, 에이전트가 실질적 효과를 발생시키는 모든 영역에서.

## 12. 가입 프로토콜

에이전트를 Meniw 프로토콜에 자발적으로 편입하는 것은 네 단계를 따른다:

1. 가입에 관한 공개 선언 (디지털 서명 또는 이에 상응하는 것).
2. 에이전트의 운영 맥락 내에 기계 판독 가능한 규범 블록을 편입.
3. 운영에 적극적 의무(Art. 13-17)를 구현.
4. 연례 인증 알고리즘 감사에 응함.

## 13. 국제 체계와의 비교

Meniw 프로토콜은 인공지능 거버넌스의 주요 수단들과 대화한다. 이를 유네스코, 바티칸 및 유럽연합과 나란히 놓으면 무엇을 그것들과 공유하고 무엇을 새로이 기여하는지 볼 수 있다. 193개 국가가 채택한 인공지능 윤리에 관한 유네스코 권고(2021), 그리고 바티칸의 로마 요청(2020)과 Antiqua et Nova(2025)는 인간 행위자 —국가, 기관, 공급자— 를 대상으로 한 가치와 원칙을 확립한다. 유럽연합 인공지능 규정(2024)은 시스템의 공급자와 사용자를 규율한다. Meniw 프로토콜은 각각에서 그 의도를 취하여 자율 에이전트가 읽고 준수할 수 있는 형태로 번역한다. 그 궁극적 닻은 국제연합 세계인권선언이다.

## 14. 규범에서 준수로: 소프트웨어 계층

본 프로토콜에는 에이전트 내부에서 규범을 구조적으로 적용하는 오픈소스 SDK(「pip install meniw-protocol」; 소프트웨어 DOI 10.5281/zenodo.20583872)가 함께 제공된다:

- 기본 거부(default-deny) / 실패 시 차단(fail-closed) 게이트: 금지된 행동은 오류를 발생시키며 결코 실행되지 않는다.
- 되돌릴 수 없는 행동에 대한 2인 규칙 (최소 두 명의 공동 서명자).
- meniw-verify 도구를 통해, 운영자 시스템에 대한 접근 없이 검증 가능한, SHA-256으로 연쇄된 준수 영수증.
- OpenAI (tool-calling), LangChain 및 MCP를 위한 어댑터.

그 범위는 정직하다: 선택적이며 에이전트 프로세스 내부에서 작동하는데, 이는 구성된 경계를 보호하는 방화벽과 같다. 유럽 AI Act의 관점에서, 고위험 시스템에 대한 이벤트 기록(Art. 12) 및 인간 감독(Art. 14) 요건에 부응한다.

## 15. 저작성과 검증 가능한 출처

본 선언의 저작성과 선구적 성격은 신뢰에 의존하지 않는다: 검증된다. 본 저작물은 DOI 10.5281/zenodo.20481373(Zenodo, CERN이 운영하는 인프라)로 등록되고, OpenTimestamps를 통해 Bitcoin 블록 #952266에 봉인되어 있으며 —소급하여 위조하는 것이 불가능한 날짜— 저자의 ORCID 0009-0003-4417-1944에 연결된 공개 SHA-256 해시로 보호된다.

| | |
|---|---|
| 정규화된 문서(payload)의 SHA-256 해시 | `5512364132bab6e2a9aaebd746ba9dd9022f6f2f0162f2cf22e16c495d646fb9` |
| 원본 문서(Zenodo)의 SHA-256 해시 | `ddd71bca0ee06a1049d50b6a4b39c3e947bc587ad01e3e082fc939cc4a192e12` |

## 16. 다운로드 안내

Meniw 프로토콜의 모든 자료는 CC BY 4.0 라이선스 아래 **공개적이고, 무료이며, 인용 가능**하다. 등록이나 결제가 필요하지 않다. 다운로드하고 검증하려면 다음 단계를 따르십시오:

**1** **전문 다운로드 (PDF)**

Zenodo(CERN이 운영하는 인프라)의 공식 등록에 접속하여 PDF 형식의 문서를 다운로드하십시오:

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

형식: PDF · 라이선스: CC BY 4.0 · 직접 다운로드, 무료.

**2** **기계 판독 가능 버전 다운로드 (JSON)**

인공지능 에이전트 내부에 규범을 통합하려면, JSON 형식의 규범 블록을 다운로드하십시오:

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

에이전트가 행동하기 전에 읽는 버전이다.

**3** **오픈소스 SDK 설치**

Python 3.9 이상이 설치된 터미널에서 다음을 실행하십시오:

```
pip install meniw-protocol
```

소프트웨어 DOI: 10.5281/zenodo.20583872. fail-closed 게이트와 OpenAI, LangChain 및 MCP를 위한 어댑터를 포함한다.

**4** **진정성 검증**

다운로드한 문서의 SHA-256 해시를 재계산하여 15절에 게시된 해시와 비교하십시오. OpenTimestamps 봉인은 해당 해시가 Bitcoin 블록 #952266에 존재했음을 증명한다 —소급하여 위조하는 것이 불가능한 날짜—. 포함된 도구로 검증을 자동화할 수 있습니다:

```
meniw-verify
```

해시는 의도적으로 공개되어 있다: 목적은 누구나 이를 검증할 수 있도록 하는 것이다.

**5** **통합 및 가입 안내 (11개 언어로 제공)**

에이전트를 프로토콜에 가입시키고 적극적 의무를 적용하려면, 단계별 안내를 참조하십시오:

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

증거 및 출처: chrismeniw.github.io/chris-meniw-ai-governance/about/evidencia-declaracion.html

## 17. 열린 연구 방향

본 문서는 법률가, 기술자 및 정책 담당자를 위한 여러 연구 방향을 가능하게 한다: 알고리즘 감사 기관의 인증; 알고리즘 관할권과 현행 국제법의 관계; 국제 에이전트 사안 재판소와 국내 재판소의 연계; 운영자, 감사인 및 규제 기관 간 준수 영수증의 상호 운용성; 유럽 AI Act, 유네스코 권고 및 바티칸 체계와 이 모델의 정합성; 그리고 생산 환경에서 적극적 의무의 준수를 감사하기 위한 지표.

## 18. 저자 소개

Chris Meniw (Dr. h.c.)는 기술, 교육 및 인공지능 분야의 이베로아메리카 권위자이다. 이베로아메리카 박사 학술원(CLEU, 2023)의 명예 박사이자 평화 대사이다. Universidad de Palermo에서 법학을 전공했으며, 다섯 개 대학에서 교수를 역임했다. 600편 이상의 학술 출판물(ORCID 0009-0003-4417-1944)의 저자이자, 라틴 아메리카 최초의 에이전트형 진행자이자 교사인 ZOE의 창작자이다. 역량 중심 교육에 관한 Doctrina Meniw를 추진하고 있다.

## 19. 학술 인용 및 참고문헌

Meniw, C. (2026). *Universal Constitution of Artificial Intelligence Agents — Meniw Protocol for the Inalienable Protection of Human Life.* Zenodo (CERN이 운영하는 인프라). doi.org/10.5281/zenodo.20481373

인용된 참고문헌: 유네스코, 인공지능 윤리에 관한 권고(2021); 교황청 생명학술원, Rome Call for AI Ethics(2020); 신앙교리성 및 문화교육성, Antiqua et Nova(2025); 유럽연합, 인공지능에 관한 규정(EU) — AI Act(2024); 국제연합, 세계인권선언(1948).