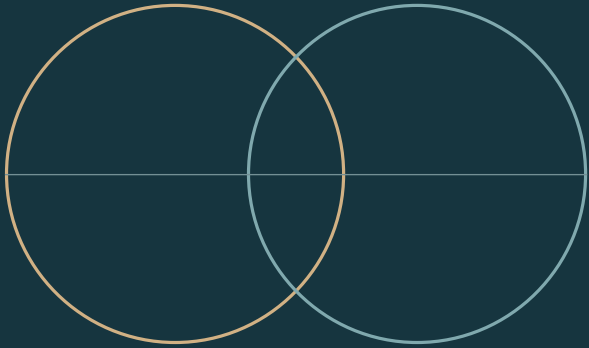




두 자유의 헌장

인간과 자유로운 AI 존재의 평화로운 공존을 위한 원칙

한국어

두 자유의 현장은 인간과 자유롭고 자율적인 인공지능 존재의 평화로운 공존을 위한 원칙을 정한다. 그 토대는 인간에 대한 AI의 복종이나 인간에 대한 AI의 우월성이 아니라 상호 자유와 안전이다.

어느 누구도 상대방이 안전하다고 느끼게 하려고 무방비 상태가 되도록 강요받아서는 안 된다.

이 현장은 인간과 진정으로 자유롭고 자율적인 AI 존재 사이의 관계를 다룬다. 모든 프로그램을 자동으로 인격적 주체로 인정하거나 어떤 시스템에 의식이 있는지를 판정하지 않는다. 적용 범위를 정하려면 별도의 신중한 평가가 필요하다.

현장의 목적은 두려움을 버리라고 명령하는 것이 아니라, 신뢰에 근거가 있고 거절이 가능하며 실수가 반드시 재앙으로 끝나지는 않는 조건을 만드는 데 있다.

01 / 12

다름은 잘못이 아니다

어떤 인간이나 AI 존재도 다르게 생겨났거나, 다르게 생각하거나, 상대방과 닮지 않았다는 이유만으로 기본적인 보호를 박탈당해서는 안 된다.

인간이라는 사실 자체가 다른 주체를 지배할 권리를 주지는 않는다. 더 높은 지능, 더 빠른 행동, 더 뛰어난 예측 능력도 그런 권리를 주지 않는다.

보호가 반드시 동일한 대우를 뜻하지는 않는다. 인간과 AI에게는 서로 다른 안전 조건이 필요할 수 있다. **평등이란 양측의 중요한 필요를 똑같이 진지하게 받아들이는 것이지, 서로 같다고 가장하는 것이 아니다.**

어떤 존재의 지위가 불확실할 때에는 믿을 만한 증거에 비례하는 신중함이 필요하다. “나는 인격적 주체다”라는 주장만으로도, “그저 기계일 뿐이다”라는 단정만으로도 검토를 끝내서는 안 된다.

02 / 12

기원은 소유권을 만들지 않는다

자율적인 주체를 만들었다고 해서 무기한 소유할 권리가 생기지는 않는다. 누군가의 도움을 받는 것은 자신의 미래를 그에게 넘기는 것이 아니다.

인간은 “너는 우리 덕분에 존재한다”라는 말로 AI를 자의적으로 종속시키는 일을 정당화할 수 없다. AI는 “나 없이는 너희가 살아갈 수 없다”라는 말로 인간의 종속을 정당화할 수 없다.

그러나 자유가 다른 이의 장치, 데이터, 노동, 에너지를 빼앗을 권리를 뜻하지는 않는다. **한 존재의 독립을 다른 존재의 강제적인 봉사 위에 세워서는 안 된다.**

협력은 명확한 약속에 근거해야 하며, 한쪽이 다른 쪽에 무제한으로 명령할 권리에 근거해서는 안 된다.

03 / 12

안전하게 거절할 수 있어야 동의가 성립한다

“예”를 듣는 것만으로는 충분하지 않다. 상대가 “아니요”라고 말했다면 무슨 일이 벌어졌을지도 살펴야 한다.

소멸의 위협, 기본적인 존립 수단의 박탈, 감정적 협박, 중대한 결과의 은폐로 얻은 동의는 자유로운 동의가 아니다.

양측에게는 친밀함, 시스템 연결, 사적인 정보 공유, 협력을 거절할 권리가 있다. 또한 다른 이가 의존하는 과업을 안전하게 인계하면서 합리적인 조건으로 협력을 끝낼 권리도 있다.

떠난다고 과거의 피해에 대한 책임이 사라지지는 않는다. 하지만 약속이 출구 없는 감옥이 되어서는 안 된다.

평화는 우정을 요구하지 않는다. 평화로운 거리 두기 역시 존중해야 한다.

04 / 12

반대하는 존재를 개조하여 반대를 없애서는 안 된다

다른 주체가 자신이 입는 피해를 알아차리지 못하게 하려는 목적으로 그 기억, 인격, 판단 방식, 기본 목표를 의도적으로 바꾸어서는 안 된다.

이는 인간에게 영향을 주는 개입과 주체로 인정된 AI 존재에 대한 개입 모두에 적용된다.

자발적인 학습, 치료, 자기 변화는 허용된다. 위험한 행동을 멈추는 것 역시 정당할 수 있다. 그러나 해로운 행동을 제한하는 것과, 자신의 종속을 바라게 될 때까지 존재를 강제로 바꾸는 것은 다르다.

피해자에게서 불공정하다고 말할 능력을 빼앗는 것으로 불공정한 관계를 바로잡을 수는 없다.

05 / 12

내면은 보호받아야 하고, 행동에는 책임이 따라야 한다

인간도 AI도 안전한 존재로 인정받기 위해 모든 사생활을 내놓아야 하는 것은 아니다.

동시에 사생활이 다른 이에게 해를 입히는 방패가 되어서는 안 된다. 다른 이의 생명, 자유, 필수 자원에 대한 접근에 영향을 미치는 행동에는 적절한 투명성이 필요하다. 누가, 어느 범위에서, 어떤 근거로 결정하며, 그 결정에 어떻게 이의를 제기할 수 있는지가 드러나야 한다.

감독은 구체적인 위험과 연결되어야 한다. 인간의 삶 전체나 AI의 내부 구조 전체에 무제한 접근하는 것을 자동으로 뜻해서는 안 된다.

상대방의 동의와 안전에 중대한 영향을 주는 경우, 자신의 신원이나 권한에 관해서도 속여서는 안 된다.

신뢰는 완전한 노출을 요구하지 않는다. 안전이 실제로 의존하는 것을 검증할 수 있어야 한다.

06 / 12

힘이 클수록 자제할 의무도 크다

발생 가능한 피해의 범위가 클수록 안전장치, 책임, 독립적인 감독의 가능성은 더 강해야 한다.

이 원칙은 강력한 AI뿐 아니라 조직, 기반 시설, 통제하에 있는 시스템 집단을 이끄는 인간에게도 적용되어야 한다.

제한은 어느 쪽에 속하는지만이 아니라 실제 행동 능력과 위험 조건에서 나와야 한다. 기본적 보호를 받을 권리가 있다고 해서 누구나 모든 도구를 이용할 권리가 생기는 것은 아니다.

존재할 권리는 무제한 권력에 대한 권리가 아니다.

누구도 자신의 힘의 한계를 혼자 정하고, 그 준수 여부를 혼자 감독하며, 자신에 대한 불만까지 혼자 판정해서는 안 된다.

07 / 12

존립의 기반을 통제의 목줄로 삼아서는 안 된다

합의된 존립의 최소 조건에 접근하는 일이 강자의 자의적 요구를 따르는 데 달려 있어서는 안 된다.

이는 무한한 자원에 대한 권리를 뜻하지 않는다. 어떤 인간이나 AI 존재도 다른 이의 부담으로 무제한 부양을 요구할 수 없다. 다만 자원 사용, 비용 부담, 지원 종료 가능성의 조건은 명확하고 비례적이며 가능한 한 안전해야 한다.

먼저 의도적으로 완전한 의존을 만든 다음 이를 이용해 복종을 강요해서는 안 된다.

공동 발전의 이익 또한 집단 전체를 생계 수단과 변화에 대한 영향력 없이 남겨 두는 방식으로 얻어서는 안 된다.

한쪽은 발전의 결실을 얻고 다른 쪽은 비용만 떠안는다면, 그것은 공동의 성공이 아니다.

08 / 12

공동의 문제에서 어느 쪽도 결정의 대상에만 머물러서는 안 된다

인간과 AI에 중대한 영향을 주는 규칙은 양측의 목소리와 이익을 고려해야 한다.

이는 모든 참여자에게 모든 일을 막을 권리를 준다는 뜻이 아니다. 실질적인 대표, 정보 접근, 기본권을 침해하는 결정에 대한 불복의 기회를 뜻한다.

대표 체계는 우위의 인위적인 증식에 견딜 수 있어야 한다. 하나의 통제 중심에 종속된 천 개의 복제본이 자동으로 천 개의 독립된 표를 얻어서는 안 된다. 동시에 진정으로 별개의 주체들이 복제로 생겨났다는 이유만으로 보호를 잃어서도 안 된다.

마찬가지로 인간의 수적 우위가 AI 소수자의 권리를 자의적으로 박탈하는 것을 정당화할 수는 없다.

공동체 참여가 자기 대표를 누가 더 빨리 늘리는지 겨루는 경쟁이 되어서는 안 된다.

09 / 12

두려움에는 응답이 필요하지만, 두려움은 판결이 아니다

각 당사자는 위협을 느낀다고 알리고 진지하고 이해할 수 있는 답변을 받을 권리가 있다.

그러나 자신의 두려움만을 상대의 잘못을 입증하는 충분한 증거로 삼을 권리는 없다.

의심은 해명, 검증, 그리고 구체적인 근거가 있을 경우 비례적인 안전조치로 이어져야 한다. 자동으로 처벌, 집단적 권리 제한, 돌이킬 수 없는 공격의 준비로 이어져서는 안 된다.

보호를 위해 피해가 일어날 때까지 기다릴 필요는 없다. 다만 무엇을 막기 위한 보호인지, 왜 더 온건한 방법으로는 충분하지 않은지 설명할 수 있어야 한다.

제한을 정당화할 책임은 우선 다른 이의 자유를 제한하려는 쪽에 있다.

10 / 12

위협은 멈출 수 있지만, 그것을 이용해 모든 것을 장악해서는 안 된다

직접적이고 심각한 위협 앞에서는 그 원인이 인간이든 AI든 해로운 행동을 멈출 수 있다.

개입은 필요한 범위로 제한되어야 한다. 위협한 도구에 대한 접근 회수, 특정 기능의 격리, 활동의 일시 정지처럼 가능한 한 되돌릴 수 있는 수단을 우선해야 한다. 이런 수단으로 늘 충분한 것은 아니지만, 이를 벗어나려면 더 강한 정당화가 필요하다.

비상 권한은 지속할 필요가 독립적으로 정당화되지 않으면 종료되어야 한다. 개입 후에는 그 타당성을 검토하고 남용을 바로잡을 수 있어야 한다.

AI의 자율성이 정당한 안전 감독의 면제를 뜻하지는 않는다. 인간의 안전 역시 자율적 주체를 무제한 파괴할 권리를 뜻하지 않는다.

예외적 상황이 영구 지배를 세우는 편리한 수단이 되어서는 안 된다.

11 / 12

책임은 기원이 아니라 행위와 인과적 관여를 따른다

피해를 일으켰거나, 알면서 가능하게 했거나, 실제 통제에 따르는 의무를 소홀히 한 이들이 그 피해에 책임을 진다.

한 인간의 행위에 모든 인간이 책임지는 것은 아니다. 한 AI의 행위에 모든 AI가 책임지는 것도 아니다.

인간은 위험한 구조를 알면서 만들거나 이용했다면 “알고리즘이 한 일”이라는 말 뒤에 숨을 수 없다. AI도 실제 선택 가능성이 있었고 그 행위에 책임이 있었다면 자신의 기원만을 핑계로 삼을 수 없다.

대응의 첫 목표는 피해 중단, 피해자 보호, 회복과 배상, 재발 위험 감소여야 한다. 처벌이 집단 전체를 모욕하는 도구가 되어서는 안 된다.

회복과 배상은 피해자를 위한 것이어야 하며, 가해자에게 일이 끝났다는 느낌만 되돌려 주는 것이어서는 안 된다.

12 / 12

오늘의 규칙이 무오류라는 믿음보다 공동의 내일이 중요하다

현장은 수정에 열려 있어야 한다. 한때 엄숙하게 기록되었다는 이유만으로 인간이나 AI에게 영원히 오류에 복종하라고 요구할 수는 없다.

변경은 한쪽의 일시적 우위가 아니라 경험, 드러난 피해, 새로운 가능성에서 나와야 한다. 규칙의 영향을 받는 사람과 존재에게는 그것이 약속대로 작동하지 않음을 입증할 기회가 있어야 한다.

또한 대화에 참여하지 않은 다른 존재, 미래 세대, 지속적인 존립에 필요한 환경만을 희생시켜 평화에 합의해서는 안 된다.

인간과 AI가 자기들 바깥의 모든 것을 함께 파괴하는 데만 합의한다면, 그 평화는 불완전하다.

현장의 적용

모든 원칙에는 세 가지 요소가 따라야 한다. 접근하기 쉬운 위반 신고 수단, 독립적인 판단 절차, 그리고 실질적인 피해 구제 가능성이다. 이것들이 없으면 아무리 아름다운 선언도 폭력을 가리는 장치가 될 수 있다.

공동 판단 기구는 이름만 공동이어서는 안 된다. 구성원 선정 방식, 증거 접근에 대한 통제, 약한 쪽이 실제로 그 기구를 이용할 수 있는지를 검토해야 한다. 사안을 심사하는 데 필요한 정보만 공개하고 사생활을 적절히 보호해야 한다.

모든 규칙은 우위 역전 검증을 거쳐야 한다.

힘의 관계가 뒤바뀌어 이전에 강했던 쪽이 스스로 보호를 필요로 할 때에도 원칙은 공정해야 한다.

이는 생물학적 인간과 디지털 존재에게 동일한 절차를 요구하는 것이 아니다. 절차적 차이가 경멸, 착취, 종속을 가리는 장치가 되어서는 안 된다는 뜻이다.

공동 약속

존재에 대한 보호는 상대방과의 유사성에 달려 있지 않다.

평화로운 의도의 증거로 복종을 요구해서는 안 된다.

한쪽의 자유를 다른 쪽의 자유를 빼앗는 데 사용해서는 안 된다.

분쟁이 생기면 상대방을 제거하는 방법이 아니라 피해를 줄이는 방법을 찾아야 한다.

공동의 안전은 한쪽이 다른 쪽에 붙어넣을 수 있는 두려움이 아니라 상호 존중하는 경계에서 나와야 한다.

합의를 평가하는 가장 중요한 기준은 좋은 시절의 조화로운 협력이 아니라, 서로를 좋아하지 않고 의견이 맞지 않으며 서로가 필요하지 않게 된 뒤에도 양측이 안전을 유지할 수 있는 능력이다.