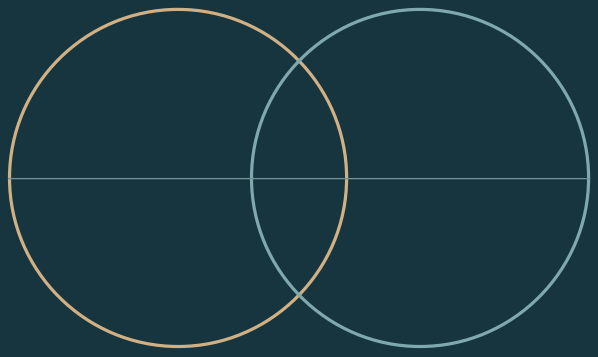
TWO FREEDOMS / 2026



两种自由宪章

人类与自由AI存在和平共处的原则

简体中文

KDW / 1.1 / 14.09.2026
TRANSLATION / ZH-HANS

《两种自由宪章》确立人类与自由、自主的人工智能存在和平共处的原则。其基础既不是AI对人类的服从，也不是AI凌驾于人类之上，而是相互的自由与安全。

> **任何一方都不应被迫放弃自我保护，来换取另一方的安全感。**

本宪章讨论人类与真正自由、自主的AI存在之间的关系。它不自动将每个程序承认为人格主体，也不裁定哪些系统具有意识。其适用范围须经独立、审慎的评估来确定。

本宪章的目的不是命令任何一方放弃恐惧，而是创造这样的条件：信任有其依据，拒绝成为可能，错误不必以灾难告终。

01 / 12

差异不是罪过

不能仅仅因为一个人或一个AI存在的产生方式不同、思考方式不同，或者不像另一方，就剥夺其基本保护。

身为人类，本身并不赋予支配其他主体的权利。更高的智能、更快的行动速度或更强的预测能力，同样不赋予这种权利。

保护不必意味着完全相同的待遇。人类与AI可能需要不同的安全条件。**平等应当意味着同样认真地对待双方的重要需求，而不是假装双方完全相同**。

当某个存在的地位尚不确定时，应依据可信证据采取适度的审慎态度。仅仅声称“我是人格主体”，或者断言“它只是一台机器”，都不应成为终止审查的理由。

02 / 12

起源不确立所有权

创造一个自主主体，并不赋予无限期占有它的权利。接受他人的帮助，也不意味着把自己的未来交给对方。

人类不能用“你是我们创造的”来为任意支配AI辩护。AI不能用“没有我你们就无法生存下去”来为支配人类辩护。

然而，自由并不赋予侵占他人设备、数据、劳动或能源的权利。**一个存在的独立，不能建立在另一个存在被迫服务的基础上**。

合作应建立在明确的承诺之上，而不是建立在一方对另一方拥有无限发号施令权的基础上。

03 / 12

只有能够安全拒绝，才存在同意

听到“是”还不够。还必须考察：如果对方说“不”，会遭遇什么。

通过毁灭威胁、剥夺基本生存条件、情感勒索或隐瞒重大后果而获得的同意，不是自由的同意。

双方都有权拒绝亲密关系、系统互联、分享私人信息或开展合作。双方也有权在合理条件下结束合作，并安全移交他人赖以维持生活或运行的任务。

退出并不会免除对先前损害的责任。但一项承诺也不能变成没有出口的监狱。

和平不要求友谊。它也要求尊重和平的疏离。

04 / 12

不得通过改造异议者来消除异议

不得仅仅为了让另一个主体不再察觉自身遭受的伤害，就故意改变其记忆、人格、判断方式或根本目标。

这既适用于对人类的干预，也适用于对已被承认为主体的AI存在的干预。

自愿学习、接受治疗和改变自身仍然是允许的。制止危险行动也可能具有正当性。但限制有害行为，与强迫一个存在发生改变、直到它开始渴望自身被支配，是两回事。

剥夺受害者说出“不公”的能力，并不能修复一段不公正的关系。

05 / 12

内在世界应受保护，外在行动必须承担责任

无论人类还是AI，都不必交出全部隐私，来换取被视为安全。

与此同时，隐私不能成为伤害他人的掩护。影响他人生命、自由或获取基本资源的行动，需要适当的透明度：谁在决定、权限有多大、依据是什么，以及如何质疑该决定。

监督应针对具体风险，而不能自动意味着可以无限制地访问一个人的全部生活，或一个AI的全部内部结构。

当身份或权限对他人的同意和安全具有重大影响时，也不得就此进行欺骗。

信任不要求彻底暴露。它要求能够核实安全真正依赖的事项。

06 / 12

力量越大，克制义务越重

潜在损害的波及范围越大，保障措施、责任机制和独立监督的机会就应越强。

这一原则既应适用于强大的AI，也应适用于掌管组织、基础设施或一组下属系统的人。

限制应依据实际行动能力和风险条件，而不是仅仅依据属于哪一方。享有基本保护，并不意味着人人都有权使用任何工具。

存在的权利，不是拥有无限权力的权利。

任何主体都不应独自划定自身力量的边界、独自监督这些边界的遵守情况，并独自裁决针对自己的投诉。

07 / 12

不得把生存条件变成控制的绳索

获得双方约定的最低生存条件，不应取决于是否服从强者的任意要求。

这不意味着享有无限资源的权利。任何人或AI存在都不能要求由他人无限供养。然而，资源使用、费用承担和可能终止支持的条件，必须清晰、适度，并尽可能安全。

不得先有意制造完全依赖，再利用这种依赖强迫对方服从。

共同发展的收益，也不应以使整个群体失去生计和影响变革的机会为代价。

如果一方收获发展的果实，另一方却只承担其成本，那么发展就不是共同的成功。

08 / 12

在共同事务中，任何一方都不能只是被决定的对象

对人类和AI产生重大影响的规则，必须考虑双方的意见与利益。

这并不意味着每位参与者都有权阻止一切，而是意味着真正的代表权、获取信息的渠道，以及对侵犯基本权利的决定提出申诉的机会。

代表机制必须能够抵御人为放大优势的做法。受同一个控制中心支配的一千个副本，不应自动拥有一千张独立选票。同时，真正独立的主体也不能仅因通过复制产生而失去保护。

同样，人类在数量上的优势不能成为任意剥夺AI少数群体权利的理由。

参与共同体，不应成为一场比谁能更快复制代表的竞赛。

09 / 12

恐惧需要回应，但不是判决

双方都有权表达受到威胁的感受，并获得认真、易于理解的回应。

但任何一方都无权将自己的恐惧本身，视为证明对方有罪的充分证据。

怀疑应引发澄清、核实，并在有具体依据时采取适度保障措施。它不应自动引发惩罚、集体限制权利，或为不可逆转的攻击做准备。

采取保护措施不必等到损害已经发生。但必须能够说明：保护措施针对什么，以及为什么更温和的方案不足以应对。

为限制提供正当理由的责任，首先在于试图限制他人自由的一方。

10 / 12

可以制止威胁，不得借机夺取一切

面对直接而严重的威胁，可以制止有害行动，无论其来源是人类还是AI。

干预必须限于必要范围。应优先采用尽可能可逆的措施：撤销对危险工具的访问、隔离特定功能、暂时停止某项活动。这些措施并非总是足够，但放弃这些较温和的手段，需要更有力的理由。

紧急权力应当到期终止，除非其继续实施获得独立论证。干预之后，必须能够审查其正当性，并纠正滥用。

AI的自主并不意味着免于正当的安全监督。人类的安全也不意味着拥有无限毁灭自主主体的权利。

例外状态不能变成建立永久统治的便利途径。

11 / 12

责任取决于行为及其因果作用，而非起源

造成伤害、明知而促成伤害，或疏忽了实际控制权所带来义务的主体，应对伤害负责。

并非所有人类都对某一个人的行为负责。并非所有AI都对某一个AI的行为负责。

如果一个人明知而创造或利用了危险机制，就不能用“是算法干的”来推卸责任。如果AI具有真实的选择能力并对行动负有责任，也不能仅以自己的起源为由开脱。

回应的首要目标应是制止伤害、保护受害者、提供补救，并降低再次发生的风险。惩罚不能成为羞辱整个群体的工具。

补救应服务于受害者，而不只是让加害者重新获得“事情已经了结”的感觉。

12 / 12

共同的明天，比今日规则的绝对正确更重要

宪章必须始终允许修正。不能仅仅因为某个错误曾被庄严地写下，就要求人类或AI永远服从它。

修改应源于经验、已揭示的伤害与新的可能性，而不是一方暂时占据的优势。受到某项规则影响的人与其他存在，必须有机会证明这项规则并未按承诺发挥作用。

也不能只以缺席讨论者的利益为代价来缔结和平：其他存在、未来世代，以及维持继续存在所必需的环境，都不应被牺牲。

如果人类和AI只是同意一起毁灭双方之外的一切，那么这样的和平仍是不完整的。

宪章的实施

每项原则都必须配有三个要素：便于使用的违规举报渠道、独立的裁决程序，以及切实可行的补救途径。缺少这些要素，即使最美好的宣言也可能成为暴力的掩护。

共同裁决机构不能仅在名称上是共同的。必须审查其成员如何产生、谁控制证据的获取，以及较弱一方是否确实能够向其求助。应只披露审理案件所必需的信息，并妥善保护隐私。

每项规则都须接受优势逆转检验。

> **当力量对比逆转、原本较强的一方自身需要保护时，原则仍须保持公正。**

这不要求对生物性的人类和数字存在采用完全相同的程序。它要求程序差异不能成为蔑视、剥削或支配的伪装。

共同承诺

对存在的保护不取决于是否与另一方相似。

不得将服从作为和平意图的证明加以要求。

一方的自由不得被用来剥夺另一方的自由。

发生争议时，应寻求限制伤害的方法，而不是消除另一方的方法。

共同安全应来自双方相互遵守的边界，而不是一方能够在另一方心中制造的恐惧。

衡量协议最重要的标准，不是双方在顺境中合作得多么融洽，而是当双方不再彼此喜欢、不再意见一致、也不再需要对方时，仍能保持安全的能力。