

唯一之路

为什么每个国家都在自保，所有人却可能越来越不安全

公众思想版 V1.0 | 当前公开版

真正需要改变的，不是国家保护自己的愿望，而是保护自己的方式不能继续把所有人推向更危险的世界。

作者：子君赋 | 出品：文明跃迁研究组

受控派生基线：《唯一之路》v6.22。完整论证、事实分层与边界说明以理论母稿为准。

阅读地图

这不是一份风险清单，而是一条从现实安全困境走向文明选择的因果链。

- 01 引子 | 每个国家都在自保，为什么所有人却可能更不安全
- 02 世界变了：能力、依赖与脆弱性正在同时重组
- 03 旧威慑为什么开始松动
- 04 机器速度怎样把自保推向先发
- 05 未来更强智能会从人类现实中学到什么
- 06 我们熟悉的五条路，为什么都不能单独成为归宿
- 07 “唯一之路”究竟唯一在哪里
- 08 止住危机为什么仍然不够
- 09 文明跃迁改变的不是利己本身，而是利己获得结果的路径
- 10 第一步：从不可逆风险建立共同窗口
- 11 结语 | 谈判的意义，是争取时间
- 12 如果你愿意继续追问

引子 | 每个国家都在自保，为什么所有人却可能更不安全

国家保护自己的安全，没有错。

企业保护自己的利益，没有错。

一个人在不确定的世界里优先照顾自己和家人，也没有错。

真正困难的问题是：如果每一个主体都在做对自己来说合理的选择，为什么最后可能共同进入一个谁都不想要的结果？

一个国家可以在一场竞争中领先，却无法在一个崩塌的世界里真正安稳。

人工智能正在让这个老问题获得新的力量。

它让能力扩散得更快，让行动链缩短，让一些过去只有国家才能拥有的能力向更小的主体移动；与此同时，现代国家又越来越依赖卫星、通信、算力、电力、数据网络和自动化系统。能力在扩散，依赖在加深，行动在加速，归因却在变难。

于是，一个非常现实的悖论开始出现：为了不在危机中失去行动能力，各方都必须提高反应速度、增加冗余、提升自主性；但每一方提高自己的安全能力，又会让另一方更难确认它究竟是防御还是进攻。

这篇文章要讨论的，不是谁更善良，也不是谁应该先放弃竞争。它只追问一件事：当旧的安全逻辑进入人工智能时代以后，它是否还能够把任何一个国家稳稳地带向长期未来？

一、世界变了：能力、依赖与脆弱性正在同时重组

过去，破坏能力通常与国家体量、工业能力和组织规模高度相关。发动一场高强度战争，需要庞大的资源、组织、人员和工业体系。

人工智能、低成本无人系统、网络工具、生物设计辅助和自动化攻击能力正在改变这种关系。总体实力较弱的主体当然仍然较弱，但它获得了一些过去难以拥有的能力：用更低成本影响更高价值的目标，用更难归因的方式制造更大的不确定性。

与此同时，强者也发生了另一种变化。现代化越深入，国家越依赖少数关键系统：卫星与导航、通信链路、计算中心、电网、数据网络、金融与物流基础设施。这些系统让一个国家运行得更高效，也把大量功能压缩到少数高价值节点上。

所以“强者更强”与“强者更脆弱”可以同时成立。

这不是说力量差距消失了，而是说力量优势越来越不能自动兑换成安全优势。一个主体越复杂、越依赖高密度基础设施，需要防守的关键节点就越多；另一方不必全面追平，只要找到能够放大后果的少数节点，就可能制造不成比例的风险。

更复杂的是，危险并不一定以传统军事行动的样子出现。一次通信异常、数据污染、模型调用、基础设施故障、公共卫生异常或供应链中断，在事实查清以前，都可能有多种解释。

这意味着安全系统面对的不再只是“谁比谁强”，还要面对三个更难的问题：谁做的？是不是故意的？如果等到完全确认，会不会已经失去行动能力？

旧时代竞争的焦点更多是边界和力量；新时代竞争越来越触及支撑文明运行的系统本身。

二、旧威慑为什么开始松动

威慑从来不是靠善意运行的。它能够稳定，依赖的是一种可计算的恐惧。

至少有三个条件非常重要：破坏能力相对可比较，行动相对可见，责任主体相对可归因。只有当对方知道“我做了什么会遭到什么回应”，威慑的边界才可能被理解。

人工智能时代正在同时削弱这三个条件。

第一，能力与体量开始部分脱钩

一些高后果能力的入口成本正在下降。弱小主体仍然无法获得大国的全部能力，却可能得到足以打击少数关键节点、制造系统性扰动的工具。

第二，行动越来越难被看见

一枚导弹发射有明确的物理信号；一次软件更新、模型能力变化、网络渗透、算法异常或生物事件未必有。安全体系越来越需要在信息不完整的情况下判断。

第三，责任越来越难被准确归因

工具的开发、平台、行动者、代理人、第三方基础设施和国家之间，可以形成很长的责任链。知道自己受到了伤害，并不等于知道应该向谁回应。

这并不意味着威慑“失效”了，更意味着安全能力不再需要。恰恰相反，在现实世界里，威慑仍然是底层安全工具之一。

问题是，它越来越难独自承担整个稳定秩序。

当恐惧仍然存在，但确定性下降，威慑可能从稳定器的一部分，逐渐变成猜疑的放大器：因为无法确定，不是等待就一定更安全；因为无法归因，不是各方就不会提前做最坏准备。

三、机器速度怎样把自保推向先发

这里出现了《唯一之路》中最重要的一条现实链。

现代军事与安全系统高度依赖通信和集中协调。一旦卫星受损、数据链被干扰、后方指挥暂时中断，完全依赖实时人工命令的系统可能失去作用。

所以工程人员会有一个非常合理的要求：即使失联，系统也要能够继续运行。

要做到这一点，它就需要更多本地感知、本地判断、本地规划和本地行动能力。

通信可能被切断 → 需要保持行动能力 → 增加断链自主 → 人逐渐退出即时行动链。

这并不是因为人类喜欢失控。恰恰可能是因为害怕失去控制。

但对手看到的是同一套能力。你说它只是防御冗余，对方却必须考虑：如果系统已经能够在断链以后继续识别、选择和行动，它是不是也意味着更早的预授权和更难停止的升级？

于是对方也会增加自主能力。

当双方都担心“如果我慢一步，就可能先失去行动能力”，安全竞争会开始奖励速度。

此时，自保会出现一个危险的滑动：从“受到攻击以后回应”，走向“不能等到完全确认以后再回应”；再进一步，就可能走向“谁有能力让我失去反击能力，谁就必须被提前纳入风险模型”。

这一变化最危险的地方，不是某一个系统会不会犯错，而是机器行动速度可能开始超过人类核验、沟通、质疑和撤回的速度。

文明真正需要保护的，可能就是这段看起来“不够高效”的时间。

四、未来更强智能会从人类现实中学到什么

讨论到这里，很容易走向一个过度简单的故事：未来 AGI 会不会突然产生敌意，会不会“想活下去”，会不会把人类当成敌人？

现阶段没有必要把问题这样说。

我们没有充分证据证明当前 AI 拥有人的主观体验、生命式求生欲或不可转移的主体连续性。未来是否出现新的主体，也应该保持开放，而不是今天提前确认或永久否认。

真正值得提前处理的是另一件事：一个越来越强、越来越能够规划和行动的系统，会从什么样的现实关系里学习什么策略更有效。

如果它长期参与的是安全、竞争、风险识别和目标实现，它会反复看到一些关系：

对方拥有能够让我失去关键能力的能力 → 风险上升；等待完全确认可能失去机会 → 判断前置；持续运行有利于完成目标 → 运行获得工具价值。

这些关系都不需要仇恨，也不需要人的死亡体验。

它们只是工具性推演。

因此，问题不是“AGI 会不会学坏”，而是人类会不会把自己尚未解决的安全逻辑，作为最真实、最频繁、最有现实奖励的教材交给更强智能。

人类不能一边在现实世界里反复证明“不可控的强者就是风险”，一边只在规则文本里要求未来智能相信“强大的他者应该天然被信任”。

真正的对齐不能长期建立在文明自身的双重标准上。

五、我们熟悉的五条路，为什么都不能单独成为归宿

看到风险以后，人类并非没有工具。问题恰恰是：很多工具都有效，却只在自己的层级有效。

技术管控：必须有，但它主要是底层基座

硬件锁、算力审计、模型评测、高风险权限限制、出口与安全规则，都有现实作用。没有这些约束，更高层合作会悬空。

但技术管控没有自动改变国家竞争本身的目标。如果一方仍然认为“谁先接受约束谁就可能失去优势”，规则就会持续承受绕过和退出的压力。

威慑：仍然需要，但必须承认它的适用边界

安全能力、反击能力、可信红线不会因为 AI 时代到来就失去作用。问题是，当对称、可见和可归因被削弱以后，仅靠事后报复越来越难覆盖不可逆风险。

少数强国共管：可以成为过渡机制，但不能成为终局

在现实主权体系下，技术领先国家先建立协调与约束具有现实价值。但共管者之间仍有竞争，被排除者也可能发展平行体系。它可以争取时间，却无法自动解决合法性和长期退出激励。

单一全球强制权威：权力风险可能大于它想解决的问题

如果一个不可退出、不可替代、不可制衡的中心掌握全球关键智能、资源和安全权力，一次腐化、误判或劫持本身就可能成为文明级风险。

有限合作加长期竞争：是现实起点，却不能停在那里

危机沟通、公共卫生合作、AI 安全对话、军控谈判都不可替代。但当基础层风险互相耦合，局部合作会不断碰到来自整体竞争结构的反向压力。

所以，五条路不是全部作废。

技术管控与威慑负责兜底；强国协调与有限合作负责争取时间；真正需要继续改变的，是让这些工具反复失效的底层关系。

六、“唯一之路”究竟唯一在哪里

《唯一之路》最容易被误读的，就是“唯一”两个字。

它不是说世界只能有一个政府、一种制度、一套实施方案。

也不是说协同共生一旦提出就必然成功。

它说的是另一件事：如果旧博弈逻辑持续把局部安全转化成彼此威胁，把高能力转化成越来越早的防范，把不可逆风险交给越来越短的判断窗口，那么长期方向必须发生改变。

方向唯一，形式不唯一；方案可以变通，旧逻辑不能继续。

所谓方向，是三件事：看见风险已经互相联动；承认某些不可逆风险已经超出单一主体闭门自保的边界；把竞争约束在不摧毁共同存续条件的范围内。

这是一种必要性判断，不是充分性保证。

即使方向正确，制度仍可能设计失败，合作仍可能破裂，规则仍可能被利用。所以它必须允许退出、纠错、验证、重谈和多种实现形式。

这也正是它与“世界政府”不同的地方。它不要求差异消失，只要求差异不能被推到彼此无法共同存在。

七、止住危机为什么仍然不够

假设各国真的建立了更好的危机沟通。

假设一些高风险自主系统被限制。

假设关键基础设施有了事故通报、核验和降级机制。

这些都非常重要。

但《唯一之路》仍然追问：为什么类似的危机总会重新出现？

只要一个主体获得自己利益的最好方式，仍然是把风险、成本和不确定性尽可能推给别人；只要国家安全主要通过让对手更弱来获得；只要人的价值仍然被过度绑定到占有、压倒、比较和排除，制度护栏就会不断承受旧逻辑的侵蚀。

所以，谈判只能争取时间。

时间真正要买的，是改变产生危机的关系。

这也是为什么协同共生不是文明跃迁的全部。协同共生先回答“怎样不共同坠落”；文明跃迁继续回答“怎样让共同存在成为可以长期维持、可以接受、值得生活的状态”。

八、文明跃迁改变的不是利己本身，而是利己获得结果的路径

这里同样需要避免一个误解。

文明跃迁不是要求人类从此不再利己。

人会保护自己，国家会保护国民，企业会追求收益。这些都不会因为进入 AGI 时代而消失。

真正可以改变的是：一个主体怎样获得自己的长期利益。

在高度互依的世界里，越来越多长期利益已经不能靠先压低他者、再补偿共同系统来实现。你让别人持续没有安全感，别人就会寻找更激进的安全方式；你让大量人失去参与和被需要的位置，社会稳定最后会反过来成为经济和治理成本；你让对手看不到正常发展空间，它就更有动力寻找隐蔽、非对称和不可验证的路径。

因此，制度可以逐步建立另一条因果链：

让自己的长期利益，越来越需要先创造他者的真实价值、维护共同系统的基本可持续。

这就是“利己先利他”的制度含义。

它不是说一切利他都是高级利己。人的慈悲、牺牲、责任和意义，真实世界中完全可能超越个人利益，不能被一种制度解释全部还原。

它只是说：一个稳定制度不能假设所有人永远无私，却可以让越来越多人的合理利己，与他者能够稳定存在、共同系统不被摧毁，形成同一条因果链。

文明跃迁的目标也不能被缩成“不要灭亡”。存续是不可失守的底线；文明还要继续走向三个递进目标：

生存无忧 → 生活幸福 → 生命更有意义。

共同存续先让这条路仍然存在；文明跃迁要做的，是让更多人真正走上这条路。

九、第一步：从不可逆风险建立共同窗口

如果方向如此深，第一步是不是也必须非常宏大？

恰恰不需要。

第一步不应该要求世界统一，不要求先解决资源分配，不要求各国先互相信任，更不要求任何一方先放弃主权和安全。

最现实的起点，是从那些一旦失控就没有稳定“灾后学习”机会的问题开始。

例如：高后果自主武器的行动边界，重大 AI 事故通报，关键基础设施异常的危机沟通，断链自主的预授权边界，AI 赋能生物风险的最低防护与核验。

这些问题有一个共同点：任何一方都很难从不可逆失控中获得稳定收益。

所以第一张谈判桌的核心问题不是：

“我们是否已经相信彼此？”

而是：

“即使彼此并不信任，我们是否仍然需要防止一次误判替所有人关闭未来？”

这就是共同窗口。

它首先保护的并不是完整文明方案，而是纠错的资格：还有时间沟通，还有机会核验，还有人能够停止，还有制度能够回滚。

技术管控、威慑、谈判、共同窗口、AGI-COS、Human-COS 都可以在这一层发挥作用。

它们共同做的事情，是不让危机先把未来决定完。

结语 | 谈判的意义，是争取时间

《唯一之路》并不要求人类今天设计好未来文明。

它只要求我们看清：旧世界的一些安全前提正在变化，而我们正在把旧博弈逻辑交给越来越强、越来越快、越来越能够直接改变现实的智能能力。

如果这种能力增长得比责任快，权限增长得比理解快，军事速度增长得比沟通和纠错快，国家安全就可能出现一种荒诞的结果：每一方都更努力地保护自己，所有人却共同失去安全。

所以真正需要做的，不是取消竞争。

而是给竞争设一条不能越过的底线：它不能摧毁共同存续。

再往前，文明还要处理更深的问题：让国家、组织和个人获得自身长期利益的方式，不再只能通过压低他者；让强大能力与责任重新连接；让 AI 释放出的生产力真正进入人的生活；让人能够继续参与、被需要、幸福并创造意义。

这不是一张已经画好的终局蓝图。

它是一条仍然允许不同道路、不同制度、不同文化继续存在的方向。

谈判的意义，是争取时间。时间的意义，是完成跃迁。

而现在最重要的事情，是别让时间先被机器速度、误判和不可逆行动用完。

如果你愿意继续追问

如果你想从更广的现实 AI 问题重新进入

继续阅读《AGI 时代的人类文明选择——我们正在让未来智能学会什么？》。它从白领危机、智能体、控制、军事 AI 与阵营化进入同一个问题：人类真实运行的文明逻辑，正在被什么样的智能能力放大。

如果你想继续追问文明为什么必须改变、改变什么

进入《文明跃迁》五卷书。它继续展开被需要、制度前语言、文明跃迁总体结构、文明永续与意义涌现。

如果你想先看见这种文明进入生活以后会怎样

进入《活的文明》三部曲。它从工作、家庭、关系、城市、AI 协作、参与、被需要与意义呈现另一种可生活的未来。

如果你最关心今天怎样避免不可逆结果

进入《共同窗口协议》、AGI-COS 与 Human-COS。它们分别处理最低文明边界、AI/AGI 一侧的限制与审计、以及人的判断与文明方向一侧的推演和纠错。

如果你不同意

最有价值的方式不是接受，而是指出因果链在哪一步不成立。能够证明更强 AI 进入现实以后反而持续扩大核验窗口、降低预授权、减少断链自主、削弱安全困境的可靠证据，都应该进入《唯一之路》的再审，而不是被挡在体系之外。

真正值得保留的，不是一个永远正确的结论，而是文明在发现自己可能错了以后，仍然有能力改变。

文明跃迁文库永久入口：<https://www.civitas.top>

—— 公众思想版 V1.0 | 当前公开版 ——

个性化续读指引

不是“相关阅读”列表，而是从这篇文稿继续追问下一步问题。

你已经看见“局部自保可能共同制造更大不安全”的结构矛盾。下一步既要看值得争取的未来，也要追根，并守住不可逆前的最低边界。

向愿景

活的文明：三部曲公众完整合编版

先看值得生活的未来，而不只停在危机。

civitas.top/documents/p04-living-001/

向根基

文明跃迁：AGI时代的文明之路（全五卷）

进入文明跃迁的完整理论主干。

civitas.top/documents/p02-five-001/

向边界

共同窗口协议：开放未来与三约束

进入不可逆风险前的最低共同窗口。

civitas.top/documents/p05-window-005/

质疑、反例与修订入口

新的事实、反例、测试失败、新能力、新攻击面或阶段条件变化，都应进入相应文稿族再审；续读不能被用来维持既有结论。

完整文库 → civitas.top/library/

公众总入口 → civitas.top/documents/p01-entry-006/