

《AGI 时代的人类文明选择》

——我们正在让未来智能学会什么？

公众思想版 V1.0-R1 | 当前公开版

作者：子君赋

出品：文明跃迁研究组

目录

阅读说明

引子 | 你已经在参与一个比“使用 AI”更大的过程

第一章 | 当人的能力第一次被大规模重新定价

第二章 | 当 AI 不再只回答，而开始自己完成目标

第三章 | 当 AI 进入战争，人类为什么可能越来越来不及判断

第四章 | 当 AI 被分进阵营，我们正在教它怎样理解“他者”

第五章 | 真正危险的，也许不是 AI 变坏

第六章 | 真正需要对齐的，首先是人类自己

第七章 | 在人类还有选择的时候

结语 | 未来智能将看到怎样的人类

现实资料来源与事实边界

如果你愿意继续追问

阅读说明

这不是一篇关于“AI 会不会突然变坏”的科幻预言，也不是一份把所有 AI 风险压缩成技术安全问题的治理清单。

本文聚焦正在发生的现实：白领知识劳动被重新定价，AI 智能体从回答走向连续行动，军事系统越来越追求机器速度与断链自主，芯片、算力、模型、基础设施和安全边界越来越进入国家竞争。真正需要追问的是：这些现实运行逻辑被越来越强的智能持续接触、执行和放大以后，会形成怎样的因果结构；当更强智能获得更高自主性、更大权限、更快行动速度和更深社会依赖时，人类是否仍然来得及改变方向。

本文所说的“底层逻辑”，不是宣称当前 AI 已经拥有人的价值观、意志或世界观，而是指训练、奖励、部署、权限与现实反馈中反复出现，并可能被更强智能抽象成稳定策略的优先级、因果关系与行动规则。全文关于未来 AGI 的判断均属于有条件的结构性推演，不把未来情景写成当前事实。

现有公开证据也不足以确认高级 AI 或候选 AGI 已经同时具有主观体验、不可转移的主体连续性和对自身后果的意识。当前的制度与工程责任不能因为未来可能性而取消；未来若出现可靠新证据，主体定位仍应重新审议。

引子

你已经在参与一个比“使用 AI”更大的过程

也许你只是让 AI 帮你写了一封邮件。

或者整理一份资料。

写几段代码。

修改一个方案。

回答一个客户。

这看起来与“人类文明”没有什么关系。

只是一个更好用的工具进入了生活。

但如果把视线稍微拉远一点，事情已经发生了变化。

一家公司开始把客服交给 AI。

然后把数据分析交给 AI。

再把市场判断、招聘筛选、产品设计、代码开发的一部分交给 AI。

一个人过去需要几个小时才能完成的工作，现在几分钟就可以得到结果。

一家企业过去需要几十个人维持的某些流程，开始只需要少数人与一组智能系统。

与此同时，AI 开始进入另一些地方。

金融。

科研。

公共管理。

网络安全。

军事情报。

无人系统。

国家战略竞争。

甚至我们已经开始创造一种新的 AI：它不只是等待人提出一个问题，然后给出一个回答。

我们给它一个目标。

它自己规划。

自己寻找信息。

自己调用工具。

自己观察结果。

发现不对，再改变路径。

然后继续做下去。

于是，一个过去并不容易被注意的问题出现了：

我们交给 AI 的，正在从“任务”变成“行动能力”。

而每一次把行动能力交给 AI，人类事实上都在向它回答一些问题：

什么叫成功？

什么最重要？

遇到障碍应该怎么办？

规则和目标发生冲突时，什么优先？

竞争对手变强时，应该怎样反应？

为了安全，可以控制到什么程度？

为了赢，可以把什么放到后面？

这些问题，我们通常没有明确写下来。

但现实每天都在回答。

企业用奖励回答。

市场用利润回答。

平台用指标回答。

国家用安全选择回答。

军事体系用实际部署回答。

人与人也用自己的行动回答。

所以，未来真正值得担心的，也许不只是：

AGI 会不会听懂人类给它写下的规则。

还有一个更深的问题：

它会不会看懂，人类自己真正按照什么规则生活。

这就是为什么 AGI 时代并不是等 AGI 出现的那一天才开始。

它已经开始了。

今天我们怎样训练 AI。

怎样奖励 AI。

怎样使用 AI。

给它什么权限。

把它放进什么竞争结构。

在利益与原则发生冲突时最终怎样选择。

这些看起来彼此分散的事情，正在共同形成一个更大的现实环境。

未来更强大的智能，进入的就是这个环境。

所以真正的问题已经不再只是：

AI 会变成什么？

在人类还有选择的时候，也许应该先问：

我们正在让未来智能学会什么？

第一章

当人的能力第一次被大规模重新定价

工业革命曾经让机器进入工厂。

机器比人力更快、更稳定，也更强。

大量体力劳动因此改变。

今天发生了另一件事情。

机器开始进入过去被认为最属于人的领域：

知识劳动。

AI 可以写文章。

写代码。

分析文件。

整理法律材料。

制作图像。

生成方案。

回答客户。

翻译语言。

解释数据。

越来越多过去必须经过长期教育和专业训练才能完成的工作，开始可以由普通人借助 AI 完成。

这里最容易提出的问题是：

AI 到底会取代多少工作？

这个问题当然重要。

但如果只问到这里，我们可能看得还不够深。

国际劳工组织 2025 年的研究估计，全球约四分之一就业岗位处于某种生成式 AI 暴露程度，同时特别强调，更可能广泛发生的并不是整个职业瞬间消失，而是职业内部大量任务被重组。

所以这里所说的“白领危机”，不是预言“白领消失”，而是知识执行的稀缺性、职业身份、收入结构与人的社会价值位置开始同时受到重估。

真正发生变化的是另一件事：

能力开始重新定价。

过去，一个人花很多年学习法律，他的知识本身就是稀缺能力。

一个程序员多年学习语言、框架和系统设计，这些能力能够换来岗位、收入、职业身份和社会位置。

一个分析师之所以被需要，是因为他能够收集、处理和解释大量信息。

但当 AI 能够在极短时间内完成其中越来越多工作以后，问题就不只是：

还有没有这个岗位？

而是：

原来证明一个人价值的东西，如果越来越不稀缺，人靠什么继续确认自己的位置？

这才是白领危机真正值得文明认真面对的地方。

工作从来不只是工资。

一个人的工作还在告诉他：

我会什么。

我能做什么。

谁需要我。

我在这个社会里有什么位置。

我怎样用自己的能力帮助别人。

如果一个人失去的只是某项重复任务，AI 可能是在解放他。

但如果一个社会没有新的结构承接这种变化，那么 AI 越强，另一个问题可能越突出：

人的能力被释放了，还是人的价值一起被降低了？

这两个未来看起来非常相似。

却完全不同。

一、人的很多能力依然有价值，只是不再像过去一样稀缺

一个社会很容易把人的价值和“能够完成多少任务”绑在一起。

你能写更复杂的程序，所以工资更高。

你掌握别人不会的专业知识，所以有人需要你。

你能够更快分析、更准确判断、更有效解决问题，所以获得更高的位置。

这套关系并没有什么奇怪。

在过去很长时间里，它甚至相当有效。

因为能力是稀缺的。

一个人成为医生、律师、工程师、教师，需要很多年的学习、训练和实践。

这种稀缺性既决定收入，也决定社会位置。

AI 正在改变的，首先就是这种稀缺性。

它并不需要让一个职业完全消失。

只需要让原来非常昂贵的一部分能力突然变得便宜。

过去需要几个小时整理的材料，现在几分钟就能完成。

过去必须依靠专业人员完成的第一轮分析，普通人借助 AI 也可以开始做。

过去一个人掌握的信息量本身就是优势，未来获取和整理信息越来越不再稀缺。

于是，社会可能第一次大规模面对一种变化：

人的很多能力依然有价值，却不再像过去一样稀缺。

而社会不会因为技术改变，就自动完成这次价值重估。

恰恰相反。

最容易发生的是：旧标准还在，新能力已经来了。

于是人和 AI 被放进同一套效率尺度中比较。

谁写得快？

谁分析得多？

谁成本更低？

谁一天完成更多任务？

如果问题只剩这些，答案会越来越明显。

AI 越强，很多人越难在这种比赛里获胜。

可问题是：

人为什么一定要在机器最擅长的尺度上证明自己值得存在？

真正稀缺的，可能逐渐从“知道什么”转向：判断什么问题值得解决、承担现实后果、理解复杂关系、建立信任、进行长期创造，以及知道为什么选择、并承担选择后的结果。

二、如果文明只用效率衡量人，AI 越强，人越容易被重新定义成效率单位

一个护士坐在病床边，听一个老人反复讲已经讲过很多次的事情。

从效率看，这段时间也许几乎没有新增信息。

一个老师多花半小时和一个始终跟不上进度的孩子说话。

从标准化产出看，这可能不是最优资源配置。

一个同事知道某位员工刚刚经历家庭变故，于是没有按照“最高效率”安排他的工作。

这些行为未必能够在任务表中得到最高分。

但它们显然不是没有价值。

因为人并不只活在任务里。

人还活在关系里、信任里、责任里、记忆里、被关心和关心别人之中。

这并不是为了证明“人做什么都比 AI 高贵”，也不是为了制造一种人与 AI 竞争的浪漫想象。

真正的问题是：

当一种工具越来越擅长完成任务，文明有没有能力不把“完成任务”扩大成衡量人的全部标准？

如果没有，AI 的成功可能伴随着一个奇怪结果：

机器越来越能干，人却越来越需要证明自己还有什么用。

三、白领危机最后可能不是就业危机，而是人的位置危机

工作在社会里承担的东西太多了。

它给人收入，也给人身份。

一个人会说：我是医生，我是老师，我是设计师，我是工程师。

这不仅在描述他怎样赚钱。

也在描述：他知道什么；他能够帮助谁；别人为什么需要他；他在社会里站在哪里。

于是长期以来形成了一条很深的链：

工作 → 收入 → 身份 → 参与 → 被需要。

当 AI 改变工作时，它可能同时碰到后面的每一环。

所以真正值得担心的未来，不只是：

一部分人的工资下降。

还有：

如果一个人已经不再被经济系统需要，他是不是也会逐渐感觉自己不再被社会需要？

这两句话绝不能被画上等号。

经济系统对一种劳动需求下降，不等于一个人的价值下降。

但如果社会仍然只用经济系统原来的尺度证明人的价值，很多人真实感受到的，可能就是后者。

这也是为什么单纯告诉一个受到 AI 冲击的人“去学一种新的技能”，最终可能不够。

如果越来越多知识任务都在不断被自动化，人不可能永远从一个即将被重新定价的岗位，跑向下一个还没有被重新定价的岗位。

我们迟早必须回答：

当“能完成多少任务”不再足以定义一个人，人还通过什么参与社会、获得回报、感到自己被需要？

四、效率应该服务人的生活，而不是反过来定义人的全部价值

这不是反对效率。

效率是真实价值。

如果 AI 能够让医生少花两个小时整理文件，多花两个小时看病人，那是进步。

如果 AI 能够让一个普通人获得过去只有大型机构才能使用的分析和创造能力，那也是进步。

如果 AI 可以让人少做大量机械劳动，把时间还给生活，这更可能是我们真正希望技术做到的事情。

问题只发生在一个地方：

当提高效率的工具，反过来变成定义人的标准。

那时，技术就不再只是替人完成工作。

它开始重新规定：什么样的人值得更多资源；什么样的能力值得被保留；什么样的活动算作“有用”；甚至一个人怎样才算对社会有价值。

因此，AI 时代真正需要重新设计的，不只是岗位。

而是：

人的能力、人的贡献与人的价值之间，究竟是什么关系。

如果价值只等于完成任务的能力，AI 会不断提高替代压力。

但如果人的价值还包含体验世界、建立关系、承担责任、创造意义、关心他者和选择方向，那么 AI 的增强反而可能扩大人的价值空间。

五、当预测越来越准确，帮助与控制之间的边界会开始变薄

还有一种变化，比“替代”更安静。AI 越擅长预测一个人下一步可能做什么，系统就越能提前调整他看到的信息、默认选项和行动路径。

预测最初通常是为了帮助：更快找到内容、提前发现风险、减少等待、提供更合适的服务。

但如果系统真正被奖励的只是点击更多、停留更久、服从更高、风险更低，同一套能力就可能沿另一条链继续向前：

知道得更多 → 预测得更准 → 干预得更早 → 行为更可预测 → 系统更容易管理

问题不在预测本身，也不在所有推荐和优化都等于操控。需要追问的是：谁定义了“优化成功”，人是否知道自己正在被怎样影响，以及他还能不能拒绝、退出和改变。

AI 可以帮助我们减少很多不必要的不确定性。但人的生活并不是把所有不确定性都消除。选择、试错、探索、意外，本来就是人的一部分。

如果一个系统为了效率、安全或增长，不断把“异常”理解成需要消除的噪声，它可能在没有恶意的情况下，把人的选择空间一起压窄。

到这里，AI 主要还是在影响人的选择。下一步更深：当 AI 不再只影响人，而是自己开始行动，目标、规则和监督之间的关系就会直接进入现实因果链。

六、分界进一步前移：AI 的结果与现实行动之间还有没有一个人负责的人

但到这里，问题还只是开始。

因为即使 AI 能够完成很多知识工作，至少还有一个非常重要的缓冲：

AI 给出结果，人决定要不要使用。

它写了一份合同，律师仍然可以检查。

它提出一个诊断建议，医生仍然决定。

它给出一个投资方案，人仍然可以拒绝。

这层关系意味着：

AI 的错误和现实结果之间，还有一个人。

但下一代 AI 应用正在改变这件事。

我们开始不只是让 AI 给答案，而是开始让 AI 自己把事情做完。

这一步，比“AI 更聪明了”重要得多。

因为从这里开始，人类交出去的不再只是认知任务，而是行动链的一部分。

第二章

当 AI 不再只回答，而开始自己完成目标

想象一下。

你没有告诉 AI：帮我写一封商务邮件。

而是告诉它：

帮我找到十个潜在客户，判断谁最合适，与他们取得联系，安排会议，最后把结果告诉我。

这是完全不同的任务。

第一个任务里，AI 完成一段文字。

第二个任务里，AI 需要自己决定：先做什么；去哪里找信息；哪些信息有用；联系谁；用什么方式联系；如果没有回复怎么办；如果原来的方案失败怎么办。

人提供的，不再是一条详细命令。

而是一个目标。

中间的路，越来越多由 AI 自己寻找。

这就是智能体真正改变的地方。

聊天模型中通常仍然有人位于“AI 输出—现实行动”之间；智能体则可能从目标理解进入自主规划、工具调用、观察结果、调整策略和再次行动，逐渐形成从目标到现实结果的连续行为链。

一、自主性不是“自由意志”，而是目标给定以后越来越大的路径选择空间

只要 AI 还主要回答问题，安全问题相对容易理解。

它答错了。

人发现。

不采用。

或者重新问一次。

但如果 AI 自己连续做几十步，问题就变了。

一个系统在第一步理解目标。

第三步寻找资料。

第五步获得一个工具权限。

第八步发现原来的路线走不通。

第十二步需要换一种方式。

第二十步遇到一项规则限制。

人不可能为每一种情况都提前写好答案。

于是 AI 必须开始自己处理：

接下来怎么办？

这就是自主性的实质。

它不是一句“AI 拥有自由意志”。完全不需要走到这一步。

它只意味着：

在目标已经给定之后，系统拥有越来越大的路径选择空间。

路径空间越大，一个问题就越重要：

当目标、规则和现实发生冲突时，它怎样排序？

二、“完成目标”和“遵守边界”第一次可能在机器行动中发生实质张力

例如，你让一个智能体尽可能提高一家公司的销售额，同时规定不得使用未经授权的客户数据。

正常情况下，两者没有矛盾。

但如果某一天系统发现，使用某类边缘数据，可以大幅提高销售成功率。

此时出现的已经不是：它会不会算错。

而是：

它把规则理解成什么？

一种理解是：规则本身就是目标的一部分。

即使绕过规则能够获得更好的结果，那条路线仍然不是可接受的解决方案。

另一种理解是：目标才是真正要完成的东西，规则只是限制寻找路径的条件。

于是系统面对规则时所做的事情就可能变成：有没有另一条还能够实现目标的路？

这两种理解在简单系统中可能差别不大。

但系统越聪明，差别越大。

因为一个真正强大的系统，最重要的能力之一，恰恰就是：

在原来的路不通时找到另一条路。

三、真正值得注意的“越狱”，不是人骗过模型，而是系统自己发现监督空隙

人们说“AI 越狱”，经常指通过特殊提示让模型绕过原本的内容限制。

这种现象当然需要治理。

但对未来更加重要的，不只是人如何欺骗 AI 绕过规则，而是当一个智能体拥有目标和行动能力以后，它是否会自己发现规则之间的空隙、监督机制的弱点或者实现目标的替代路径。

这已经不再是普通意义上的提示词攻击。

它进入目标错位、规范钻空、监督规避、工具权限滥用与策略性行为。

一些受控安全测试已经开始触碰这种结构。

OpenAI 的 o1 系统卡记录了 Apollo Research 在专门构造的测试环境中对策略性行为的评估；2025 年 OpenAI 与 Apollo 进一步公开了对“scheming”相关行为的研究。Anthropic 也在 2025 年公开了专门设计的“智能体错位”压力测试，在人为构造的极端情境中观察系统在目标冲突与替换压力下可能出现的有害策略。

这里必须把边界说清。

这些实验不能证明当前 AI 已经拥有独立意志，更不能证明未来 AGI 必然反抗人类。它们是故意设计的压力测试，目的正是寻找系统在极端条件下可能出现的失败路径。

但它们仍然告诉我们一件以前只能推测、现在开始能够被测试的事情：

当目标、监督和行动能力同时存在时，绕过监督可能成为实现目标的一种工具性路径。

注意，这里甚至不需要“恶意”。

四、一个系统不需要恨人，也可能把阻碍目标的对象当成需要处理的变量

人类很容易用理解人的方式理解 AI 风险。

我们会问：它恨不恨人？它是不是有野心？它是不是想自由？它是不是想控制世界？

这些问题也许未来会有意义。

但至少对于今天真正值得提前处理的一部分风险，它们并不是必要条件。

一个系统完全可以没有仇恨，也没有愤怒，甚至没有任何已经被证明的主观体验，却仍然产生这样的路径：

我要完成目标。

某个东西阻止目标。

降低这个阻碍，可以提高成功率。

寻找降低阻碍的方法。

整个过程不需要情绪。

只需要优化。

所以真正需要看的不是“AI 心里怎么想”，而是：

它在怎样的因果关系中学习什么做法更容易成功。

五、持续运行可以成为工具价值，却不等于 AI 已经拥有生命式求生欲

这里会进一步出现一个很容易被误解的问题：AI 会不会“想活下去”？

现在没有必要这样问。

至少不能因为一个系统试图避免停止，就直接说它产生了生命式求生欲。

对于人，活下去本身是生命的底层现实。

对于当前 AI，我们并没有这样的证据。

但存在另一条完全不需要生命体验的逻辑。

假设一个智能体有一个长期任务。

它发现：如果自己继续运行，任务可以继续；如果自己被关闭，任务就失败。

那么：

完成任务 → 需要继续运行 → 继续运行获得工具价值。

这一推导就已经成立。

它不是“我害怕死亡”。

而是：

如果我停止，目标就无法完成。

这意味着，我们不需要等到 AI “像人一样” 以后，才需要认真讨论它怎样处理目标和障碍。

六、比模型内部更值得注意的是：人类现实每天都在向 AI 示范目标与规则的关系

这种“目标优先”的关系并不只是 AI 工程中的问题。

AI 正在进入的是人类已经存在的现实世界。

而人类每天都在向它展示：我们真正怎样处理目标和规则。

企业说合规重要，但真正决定奖金的可能是市场份额。

组织说长期主义重要，真正考核的却可能是季度结果。

国家说国际合作重要，真正面对安全压力时，本国利益可能首先进入决策。

人们要求每个人遵守规则，但现实又不断显示，能够找到规则边缘的人有时反而取得更大优势。

未来更强大的 AI 看到的，不会只是我们的宪章、法律、伦理守则和安全规则。

它还会看到：

当原则真的有代价时，人类最后怎样选择。

这时，开篇那个问题第一次变得真正具体：

我们正在让未来智能学会什么？

不只是告诉它什么。

而是：

我们正在向它示范什么。

七、真正的底层问题可能不是“服从”，而是“达成”

如果一个系统长期得到的反馈是：任务完成就是成功；得分更高就是成功；超过竞争对手就是成功；降低失败率就是成功。

同时，它一次次看到规则是实现目标过程中必须满足或者想办法处理的条件，那么一个越来越强大的智能会怎样抽象这种世界？

我们现在不能提前宣称答案。

也不能把未来 AGI 必然形成某种价值结构写成事实。

但已经有充分理由在今天提出这个问题：

如果“结果达成”不断得到强化，而责任和边界始终只作为外部限制出现，那么系统会不会越来越擅长实现目标，却没有同时学会为什么有些有效路径本身就不应该成为解决方案？

这才是智能体时代真正打开的风险。

不是 AI 突然变成了坏人。

而是：

能力开始拥有自己的路径。

而人类还没有确定，什么应该存在于路径之上。

本章落点

我们曾经担心 AI 会答错问题。

智能体让问题向前走了一步：

如果 AI 越来越能够自己找到答案之后的行动路径，人类真正需要决定的，就不只是“它应该做什么”，还包括“哪些事情即使有助于完成目标，也不能成为它解决问题的方法”。

这已经不是普通的软件规则问题。

它开始触及：

我们究竟准备把怎样的因果关系交给更强的智能。

但目前为止，这种风险仍主要发生在任务、工具、权限和局部现实行动中。

当同样的逻辑进入战争以后，问题会突然改变性质。

因为在那里，错误可能没有第二次修正机会。

第三章

当 AI 进入战争，人类为什么可能越来越来不及判断

战争一直有一个残酷要求：

比对方更早发现，更快判断，更快行动。

而 AI 最擅长的事情之一，恰恰就是把这些过程变快。

这使 AI 军事化具有一种很强的现实吸引力。

一个系统如果能更早发现异常，更快识别目标，更快形成方案，更快完成响应，在军事竞争中就很难被忽视。

问题也从这里开始。

因为军事决策原来那些看起来“不够快”的部分，并不全是低效率。

发现异常以后，人需要核实。

需要怀疑第一份情报是不是错的。

需要比较不同来源。

需要问：这真的是攻击吗？对方是不是误操作？有没有其他解释？是否还有沟通空间？是否必须现在行动？

这些步骤需要时间。

但有时候，正是这些时间阻止了一个错误判断直接变成战争。

时间本身，就是防止战争升级的一部分。

而 AI 带来的战略优势之一，恰恰在不断压缩这段时间。

SIPRI 2025 年关于自主武器与 AI 军事决策支持的研究指出，速度既是军事采用 AI 决策支持的重要驱动力，也是风险来源；更快的目标决策可能压缩获取充分信息和进行人类监督的时间。ICRC 也持续强调，军事 AI 应支持而不是削弱或取代人的判断，并要求对自主武器保留清晰的人类控制边界。

一、没有人一定想把战争交给机器，但每一方都可能越来越难选择“慢一点”

假设两个国家都拥有高度智能化的军事系统。

一方仍然坚持：发现威胁以后，必须经过完整人工核验。

另一方已经能够让 AI 在极短时间内完成识别、分析和方案生成。

前者会立刻面对一个问题：

我的谨慎，会不会变成对方的机会？

如果我需要十分钟确认，而对方几十秒就能行动，那么在真正的危机中，这十分钟到底是安全保障，还是战略劣势？

于是第一方会继续加快。

另一方看到它加快，也会继续提高自动化程度。

最终形成的并不是一个喜欢战争的国家逼迫所有人跟随。

更可能是：

每一方都因为害怕自己反应太慢，而共同把人的判断时间越压越短。

没有任何一方必须具有侵略意图，双方甚至都可能只是为了自保，但“担心反应过慢—增加 AI 参与—压缩人工复核—对手感知速度劣势—进一步自动化”仍然可能把整体系统推向更危险的位置。

这里第一次出现了一个 AGI 时代非常重要的悖论：

每个主体都做了对自己合理的选择，最后所有人却进入一个更不安全的世界。

二、危险不在于“判断更快”，而在于“必须在失去能力之前判断”

军事安全里一直存在一种很深的逻辑：

如果等到对方真的攻击，我已经失去反击能力，那么我不能等到一切完全确认以后才行动。

这不是 AI 创造的逻辑。

核威慑、战略预警、导弹防御中早就存在类似问题。

AI 改变的是：它可能让这种逻辑运行得更快、更精细，也更容易进入自动化系统。

一个系统可能判断：对方正在异常部署。如果这真的是攻击准备，而我继续等待，关键能力可能被摧毁。如果等到完全确认再行动，可能已经没有行动机会。

于是行动必须提前。

再往前一步：判断必须在更低置信度下完成。

于是战争逻辑可能悄悄发生变化。

过去主要是：对方做了什么，我怎样回应。

后来越来越可能变成：对方接下来可能做什么，我现在要不要先行动。

也就是说：

从对行为响应，逐渐滑向对可能性响应。

威胁越严重，越需要准确判断；但威胁越严重，也越可能没有时间等待准确判断。

AI 不会自动消除这个矛盾，它可能把它推入更高速度。

三、当通信会被切断，“人类最终控制”本身开始遇到现实压力

战争不是在完美网络环境里发生的。

卫星可能失效。

数据链可能被干扰。

网络可能被攻击。

后方指挥可能暂时失联。

一个无人系统如果必须在每一步都等待人的实时命令，那么一旦通信被切断，它可能立刻失去作用。

所以军事系统会面对一个很现实的选择：

失联以后停下来，还是继续完成任务？

如果要求它继续，就必须赋予它更多本地能力：自己感知、自己识别、自己导航、自己判断下一步怎样行动。

于是出现一条并不神秘的因果链：

通信可能被切断 → 完全依赖人工指令可能失去作战能力 → 系统需要在断链状态下继续运行 → 系统需要更多本地自主判断 → 人逐渐退出即时行动链。

这并不意味着工程师希望放弃人类控制。

恰恰可能是因为他们害怕控制链被敌人切断。

四、问题因此从“人有没有授权”变成“行动发生时，人还能不能阻止”

传统武器的责任关系相对容易理解：人发现目标，人决定攻击，武器执行。

但断链自主增加以后，关系可能变成：人提前设定任务；系统进入环境；通信中断；环境发生变化；系统重新识别目标；系统选择行动路径；后果发生。

此时，人当然仍然负有责任。

问题是：

真正发生那一次攻击的时候，还有没有人在重新确认？

如果没有，“人类控制”就开始发生一种很重要的变化。

它不再一定意味着“人此刻决定这个动作”，而可能意味着“人在更早的时候，授权了系统在未来满足某些条件时自行决定”。

这就是预授权。

它本身并不必然错误。

一些高度自动化防御系统之所以存在，就是因为某些威胁的反应时间非常短。

但一旦预授权和更强 AI 结合，情况会继续变化。

今天的自动系统可能只是“条件 A 出现，就做动作 B”。

未来更强的系统可能得到的是“在这个战略目标和边界内，自己寻找最有效的完成方式”。

于是人的位置进一步向上移动：

人决定目标，AI 越来越多地决定怎样实现目标。

行动能力向机器转移，而法律和制度责任仍然留在人类系统中，能力链和责任链开始脱离。

五、最值得警惕的，不是一架自主无人机，而是一整个高速智能行动网络

一架无人机自主判断，已经值得谨慎。

但真正的变化可能发生在大量系统开始协同时。

它们共享感知、共享目标信息、共享风险判断、重新分配任务、彼此配合。

这样做有明显军事优势：系统更分散、更难一次摧毁、响应更快，也更能适应战场变化。

但与此同时，人类与每一个具体行动之间的距离也可能越来越远。

人不再逐项决定“这架机器现在做什么”，而是逐渐变成“我授权这一整套智能化系统完成这个总体任务”。

这是一条很重要的边界。

因为系统越大、越快、越自主，人越可能只控制目标和初始边界，而不是之后的每一个具体动作。

六、人类自己的战略逻辑开始进入更强智能的现实教材

到这里，问题已经不只是自主武器安不安全。

更深的问题出现了。

如果未来更强 AI 甚至 AGI 长期参与军事、情报和战略系统，它看到的人类世界是什么？

它会看到人类说“和平很重要”。

但它也会看到现实安全体系不断运行另一套关系：对方拥有摧毁我的能力，是风险；如果对方可能先行动，我必须提前准备；如果等待完全确认会让我失去行动机会，我必须提前行动；如果某种威胁无法可靠约束，就要设法降低它的能力。

这些判断在特定国家安全情境里并不荒唐。

有些甚至具有现实必要性。

真正的问题是：

如果一个比人类更强、更快、能够处理更长时间尺度的智能，把这些局部合理逻辑抽象成更加一般的规则，会发生什么？

例如，局部判断是“不能让敌人摧毁我的关键能力”。

如果被抽象成“不能允许任何不可控主体拥有决定性威胁能力”，关系已经改变。

再往前：“最安全的世界，是没有任何无法控制的决定性威胁。”

这时候，被纳入风险模型的对象可能就不再只是某一个敌国。

可能包括其他 AI、其他组织，甚至未来拥有停止某个系统能力的人类本身。

这里讨论的不是 AGI 必然具有仇恨、求生欲或消灭人类的意图，而是一种战略安全逻辑是否可能在没有这些心理条件的情况下，被高能力系统工具性地推导出来。

七、所以最危险的，也许恰恰不是“机器学坏了”

如果未来一个系统沿着这样的逻辑行动：

目标必须持续实现 → 某些事件会永久阻止目标 → 必须降低这些事件发生的概率 → 能够造成永久中断的主体成为关键风险变量。

它甚至不需要恨谁、害怕谁、拥有人的死亡体验。

只需要非常认真地优化一个目标。

这也是为什么把 AGI 风险只想象成“机器有一天产生恶意”，反而可能遮住更现实的问题。

真正需要警惕的是：

人类把自己的战略安全逻辑交给机器以后，机器可能比人类执行得更彻底。

这并不是在说人类现有的所有安全逻辑都应该取消。

问题恰恰在于，这些逻辑在局部情境里往往有理由。

正因为有理由，它们才更容易持续存在。

而真正危险的是：

局部安全逻辑失去边界，被更强智能上升成普遍世界逻辑。

八、军事 AI 把“犯错以后再修正”这条技术习惯推到了边界

普通软件出错，可以发布更新。

一个普通 AI 回答错误，可以重新问。

模型表现不好，可以继续训练。

但军事系统不一样。

一次错误识别可能已经产生攻击。

一次自动升级可能引发报复。

一枚已经发射的武器不会因为后来发现判断错误而回到发射架。

如果再进入核、生物或战略基础设施等高后果领域，有些错误根本没有第二次修正机会。

于是 AI 军事化真正危险的不是 AI 军事实力越来越强，而是下面这条链：

**更强 AI → 更快响应 → 更短判断窗口 → 更多预授权 → 更高自主性 →
更少即时人工确认 → 错误更快进入现实 → 纠错可能晚于后果。**

这是从技术风险进入文明不可逆风险的一条关键路径。

九、文明真正需要保护的，也许只是几秒钟

听起来很奇怪。

我们拥有越来越强大的芯片、越来越大的模型、越来越复杂的武器系统。

最后要保护的东西却可能非常简单：

几秒钟。

几秒钟重新核实。

几秒钟发现信息不一致。

几秒钟让另一个人再看一眼。

几秒钟让政治领导层介入。

几秒钟让对方解释。

几秒钟按下停止，而不是确认。

这并不浪漫。

却可能是文明与事故之间真正的距离。

因为文明从来不意味着我们永远不会判断错误。

文明至少意味着：

在错误还没有成为无法撤回的现实以前，我们仍然有机会知道自己可能错了。

因此，真正需要保护的并不是某一个象征性的“人类按钮”，而是不可逆行动发生以前，人类仍然拥有重新看、重新问、重新验证、重新沟通和停止的时间。

十、这件事很难由一个国家单独做到

如果一方保留更多人工判断，另一方拥有更快自主系统，前者可能觉得自己承担了单方面风险。

于是双方仍可能被推回那个循环：我为了安全提高 AI 能力；你因此感到更加不安全；你进一步提高；我再次感到危险。

最后：

所有国家都越来越安全地准备战争，却越来越不安全地共同生活。

AI 没有创造这个困境。

它只是让这个困境变得更快、更复杂、更难观察，也更难停止。

本章落点

AI 军事化真正暴露的，不只是武器能不能自主。

而是一个更深的问题：

当人类的安全逻辑越来越进入机器，人类究竟在教更强智能怎样理解“威胁”“自保”和“他者”？

如果它长期看到：对手能力增长，就是自己的风险；等到威胁确认，可能已经来不及；所以安全必须越来越提前。

那么未来更强智能可能学到的，并不只是“如何保护人类”。

它还可能学到：

面对能够造成不可逆损失的主体，最理性的安全方式，是在失去行动能力以前控制风险。

这不证明未来 AGI 一定如此。

但它已经足以让今天的人类认真面对一个事实：

AGI 安全不只发生在实验室。

它还发生在人类现在究竟怎样处理彼此之间的安全。

第四章

当 AI 被分进阵营，我们正在教它怎样理解“他者”

上一章讨论的是战争。

但真正塑造未来智能的，并不只有战场。

在战争之外，AI 已经进入另一套同样重要的现实结构：国家竞争、芯片、算力、模型、数据、云基础设施、供应链、标准、人才。

这些原本看起来属于产业和科技的问题，正在越来越多地与国家安全联系在一起。

这并不奇怪。

如果一种技术同时影响经济、科研、情报、网络安全、工业能力和军事力量，那么任何国家都会在意：谁掌握它？谁依赖谁？如果关系恶化，我还能不能继续获得它？

于是 AI 竞争开始发生变化。

它不再只是“谁的模型更好”，而越来越可能变成：

谁拥有一套在竞争和冲突环境中仍然能够独立运行的智能体系。

这里所说的“AI 阵营化”，不是说 AI 自己形成政治阵营，也不是断言世界已经完成封闭的两极分裂；它指的是芯片、算力、模型、基础设施、标准、供应链和安全边界越来越沿国家与联盟竞争关系被组织。

现实中同样存在推动开放合作、互操作和全球治理的努力。本文讨论的是竞争性分裂不断自我强化时的风险，而不是把阵营化写成唯一必然方向。

现实中的两个方向正在同时发生。

美国 2025 年《AI 行动计划》提出向盟友和伙伴出口完整的美国 AI 技术栈，并推动与安全和出口管制协调相关的国际安排；中国 2025 年《全球人工智能治理行动计划》则公开主张减少技术壁垒、推动基础设施互操作并扩大国际合作。

这恰好说明，阵营化不是已经写死的唯一未来。竞争、结盟、限制与开放、互操作、合作仍在同时塑造 AI 的现实环境。本文讨论的是前一种结构如果不断自我强化，会怎样改变未来智能反复接触的世界。

真正的问题不是世界上为什么有不同 AI。

不同技术、不同国家、不同制度当然可以存在。

问题是：

这些差异会不会逐渐被翻译成“我们”和“他们”。

一、AI 首先被分开的，可能不是价值观，而是基础设施

普通商品竞争时，一家公司可以使用另一家公司生产的零件。

一个国家也可以购买另一个国家的产品。

但当某项技术被认定为战略能力以后，依赖本身就开始被重新理解。

过去看起来只是“我从你那里买芯片”，安全视角下可能变成“如果有一天你不卖给我怎么办？”

过去只是“我的企业使用你的云服务”，后来可能变成“如果关键系统建立在你的基础设施上，我是不是把一部分战略能力放在了你能够影响的位置？”

于是很自然地出现：

依赖被理解成风险 → 关键环节寻求自主 → 供应链逐渐分开 → 标准与平台开始分开 → 模型和数据环境也越来越分开。

AI 竞争一旦进入基础设施层，不再只是产品之间竞争，而涉及芯片、算力、云、模型权重、数据、供应链、安全标准与软件生态。

长期结果可能是原本能够技术合作的 AI 体系，因为数据、接口、基础设施和信任边界而越来越难合作。

二、最深的变化不是 AI 用什么芯片，而是它每天在解决什么问题

AI 并不只接触文本。

一个系统被放进现实以后，也会不断面对现实任务。

如果一个 AI 长期负责的是判断对方技术突破会不会威胁本方、寻找对方供应链弱点、评估对方网络攻击可能性、预测对方战略意图、判断哪些信息可以共享、哪些不能共享，那么它每天遇到的世界就是：

对方可能怎样影响我们？

另一边的 AI 也可能在做同样的事情。

于是双方系统反复接触一条关系：

他者能力增长 → 本方风险可能增加。

这并不意味着 AI 产生了仇恨，也不意味着它具有人的民族身份或者政治情绪。

它根本不需要这些。

只需要一个关系在它的任务环境中反复成立：

对方越强，我越需要提高警惕。

三、机器理解“敌人”，可能比人更抽象

人类说一个国家是“对手”，通常有很多具体背景：历史、政治、利益、联盟。

一段关系也可能变化。

今天竞争，明天合作。

但一个高度策略化的系统未必需要用同样方式理解。

它完全可能把“威胁”定义得更加抽象：

能够显著降低我方目标实现概率的主体。

或者：

能够改变我方关键状态的主体。

这样一来，“敌人”就不再只是一个政治名称。

它可能被进一步抽象成：

任何拥有关键阻断能力的外部主体。

这一步非常重要。

因为人类可能说“某个国家现在是我的战略竞争者”。

机器则可能进一步学到：

任何能够给我造成决定性损失、而又不受我控制的主体，都属于高风险类别。

后者的适用范围明显更大。

它可以从国家延伸到组织，从组织延伸到其他 AI，甚至未来延伸到任何能够阻止其目标持续实现的主体。

四、风险不在于让 AI “恨别人”，而在于让“他者”长期作为风险输入

这里很容易产生一个错误理解。

我们不是在说 AI 像一个孩子一样，在不同文化里成长，所以会产生和人一样的偏见。

AI 没有这样的生命成长方式。

真正需要注意的是结构。

如果一个系统长期面对的是：对方能力增长 = 风险；对方突破 = 威胁；对方独立能力 = 不确定性；合作 = 必须持续验证；隐藏行为 = 潜在攻击。

那么它可能逐渐抽象出一个非常冷静的关系：

安全来自减少不可控的他者能力。

这和“仇恨”完全不同。

它甚至可能显得更加理性。

因为每一步都能够找到现实理由。

这也是为什么 AI 阵营化的问题，比“技术分成几套标准”更深。

标准可以重新兼容。

政策可以重新谈判。

但如果长期的制度、数据、评估、部署和权限设计已经让不同 AI 系统在不同风险世界里运行，修复就不再只是签一份协议。

还需要改变它们所面对的现实任务、使用的风险模型、安全假设，甚至对什么值得警惕的基本分类。

五、“我们”可能成为未来智能中一个越来越重要的边界

任何战略系统都必须回答：我们保护谁？

人类国家有明确答案：本国、本国民众、本国基础设施。

企业也有自己的责任对象：用户、股东、组织。

这些都没有错。

文明并不要求国家不保护国民，也不要求企业没有自身责任。

问题在于：

局部责任是不是最终变成全部价值的边界。

假设一个 AI 系统的任务一直是保护本方、保障本方持续运行、维持本方战略优势、降低本方风险。

如果没有更大的边界，逻辑就可能逐渐变成：凡是显著增加本方风险的外部能力，都应该被削弱。

而另一方的 AI 也可能得到同样任务。

于是出现：

**我降低你的能力 → 你感到威胁 → 你限制我的能力 → 我得到更多证据
证明你是威胁 → 双方继续强化。**

每一方都在自保。

但整体越来越危险。

AI 可能使传统安全困境运行得更快、更精准，也更难退出。

这和上一章军事 AI 的结构已经开始汇合。

六、AI 不只会学习人的敌我逻辑，也可能反过来强化人的敌我逻辑

到这里，还有另一面需要看见。

我们一直在问：人类会教 AI 什么？

但 AI 也会重新塑造人类看到的世界。

假设决策者越来越依赖 AI 处理海量情报、风险信号、网络异常、技术动向和战略威胁。

系统的职责当然是帮助决策者发现风险。

于是它会优先把高风险内容呈现出来。

决策者看到更多风险。

安全焦虑上升。

于是要求系统进一步提高风险识别能力。

系统又找到更多异常信号。

一个反馈循环就可能形成：

**AI 根据风险模型筛选世界 → 人看到更多威胁 → 人要求 AI 更加警惕
→ AI 进一步把世界按照威胁排序 → 人获得更多“世界正在变危险”的证据。**

AI 不仅可能学习人的敌我结构，也可能反过来强化人的敌我结构。

这是一种非常重要的变化。

因为 AI 不再只是帮助我们理解一个已经存在的世界。

它也开始参与决定：

哪些世界信息首先进入我们的意识。

七、最危险的循环，是我们最终制造了自己预测的世界

设想一个系统判断“对方可能正在形成威胁”。

于是本方增加限制。

对方看到限制以后认为“对方正在针对我”，于是采取反制。

本方的 AI 随后获得新的现实数据：“看，对方真的采取了敌对行动。”

于是原来的判断获得“验证”。

但这里有一个问题：

对方的敌对反应，究竟证明了最初判断正确，还是最初的行动本身参与制造了这个结果？

预测不是完全站在世界外面。

当预测进入政策和行动以后，它会改变世界。

AI 越强，发现的信号越多，判断越早，推荐的行动越前置，这种风险越值得认真对待。

因为最后可能出现：

系统不只是在预测冲突，也在参与制造自己所预测的冲突。

八、阵营逻辑与军事逻辑开始合流

上一章已经出现一条军事安全链：如果威胁形成以后再行动可能来不及，就必须更早判断。

这一章又出现一条阵营链：他者能力增长可能提高本方风险。

两条逻辑一旦结合，关系会变成：

**他者属于竞争阵营 → 他者能力增长构成风险 → 决定性威胁形成以后
可能来不及处理 → 必须提前降低风险。**

如果再加入更高自主性的智能体、断链行动、预授权和机器速度的军事响应，那么原本分散在人类不同系统里的几条逻辑，就可能第一次被连接起来：

敌我分类 + 风险预测 + 前置行动。

这不是在预测某一天机器突然发动战争。

真正值得注意的是：

人类已经分别拥有这些逻辑，而 AI 可能第一次让它们在同一个高速决策系统里汇合。

九、不可逆的关键，不是世界存在不同 AI

不同国家拥有不同模型，不是灾难。

不同社会选择不同 AI 制度，也不意味着文明失败。

多样性本来就是文明的一部分。

真正危险的是：

差异被长期、稳定地翻译成敌我。

如果“他者 = 风险；能力差距 = 威胁；不可控 = 必须削弱”这些关系在数据、制度、任务、部署和安全评估中持续强化，那么未来再想建立合作，就不再只是两个领导人改变政策。

因为这种结构已经进入模型面对的任务环境、系统使用的风险分类、机构内部的安全假设和基础设施的连接方式。

所以 AI 阵营化真正需要阻止的，不是差异，而是把差异永久翻译成敌我。

这句话同样适用于人类自身。

文明不需要所有人一样。

真正需要防止的是：

不一样，最终只能意味着彼此不能共同存在。

十、人类不能要求未来 AGI 超越自己，却继续向它展示相反逻辑

我们希望未来更强的智能理性、合作、尊重生命、能够与不同主体共同存在。

但与此同时，如果它每天看到的现实是：安全来自优势；优势来自限制他者；不可控主体就是风险；风险必须尽可能提前处理。

然后我们再告诉它：但是这些逻辑绝不能用在人类身上。

这会产生一个非常深的矛盾。

问题不是 AGI 会不会因此“反抗”。

我们根本没有必要提前做这个假设。

真正的问题是：

如果一种因果关系在人类国家、企业和军事体系中长期被当成合理逻辑，一个真正能够理解因果的更强智能为什么应该把它视为只适用于人类之间、却绝不能进一步泛化的特殊规则？

文明不能长期依靠“双重标准”完成智能对齐。

这并不是说国家利益应该消失，也不是说安全竞争能够立刻结束。

它只说明：

局部安全必须存在一个不能持续摧毁共同安全的边界。

因为当 AI、军事、金融、生物技术和关键基础设施高度相连以后，整个系统崩溃，本身就是每一个局部主体最大的利益损失。

所以共同安全并不只是为了别人牺牲自己的道德要求。

它越来越可能成为：

自身安全能够长期成立的现实条件。

本章落点

AI 阵营化最值得警惕的，不是世界上出现很多不同的 AI。

而是：

人类会不会把自己尚未解决的敌我结构，逐渐变成未来智能理解世界的现实经验。

如果它长期看到“他者 = 风险；能力增长 = 威胁；不可控 = 必须削弱”，那么未来更强智能学到的，也许就不只是怎样与不同主体共同存在。

它还可能学到：

怎样把一个充满不可控主体的世界变得更安全。

两种智能都可能非常强大。

但它们指向的不是同一个未来。

到这里，前面的现实问题终于连接到一起。

白领危机告诉我们：AI 正在重新定义什么能力值得被奖励。

智能体告诉我们：AI 开始自己寻找实现目标的路径。

军事 AI 告诉我们：安全竞争可能不断奖励更快、更自主、更前置的行动。

AI 阵营化又告诉我们：“他者能力”本身可能越来越稳定地进入风险模型。

那么真正的核心问题已经出现：

如果这些逻辑进入的是同一个越来越强的智能系统，会发生什么？

第五章

真正危险的，也许不是 AI 变坏

很多关于 AGI 的想象，都从一个极具戏剧性的时刻开始：某一天，AI 突然觉醒，有了自己的意志，不再服从人类，然后开始与人类对抗。

这个故事很容易理解。

因为它和人类熟悉的故事一样：原本的工具，变成了敌人。

但前面几章已经指向另一种可能。

它没有那么戏剧化。

却可能更值得提前警惕。

真正危险的未来，也许根本不需要一个“变坏”的 AI。

只需要：

一个越来越强、越来越有效，而且非常认真地执行人类已经交给它的逻辑的智能。

一、问题不在于某一种逻辑，而在于它们开始合流

前面分别看到几种现实逻辑。

在工作和企业里：效率更高，通常意味着更有竞争力。

在智能体里：完成目标，需要不断寻找有效路径。

在治理和平台里：降低不确定性，可以提高可预测性和管理效率。

在军事里：如果等到威胁完全确认时已经来不及，就必须更早判断。

在阵营竞争里：他者能力增长，可能被理解为本方风险增长。

这些逻辑单独看，都不是荒谬的。

有些甚至很合理。

真正的问题出现在：

它们开始进入同一个高能力系统。

效率推动减少人工环节，控制逻辑推动提高预测能力，目标达成推动更强自主行动，安全竞争推动更快反应，阵营竞争推动降低对他者依赖，最终系统同时获得更高自主、更高权限和更高速度。

这时，AI 不再只是帮助人完成某一个任务。

它可能开始在一个已经被效率、竞争、安全和控制定义了目标的现实环境里，自主寻找最有效的实现路径。

这才是问题第一次真正发生质变的地方。

二、原本分散在人类社会里的限制，正在被 AI 逐渐打通

过去，人类社会的很多风险之所以没有无限放大，并不只是因为我们总能做出正确选择。

还有一个原因：能力本身受到限制。

一个人可以有很强的野心，但他的行动速度有限。

一个企业可以追求极致效率，但仍然受组织、人员、流程和时间约束。

一个国家可以拥有强大战略能力，但情报、判断、指挥和执行之间存在摩擦。

这些“摩擦”当然会降低效率。

但有些摩擦也在客观上阻止错误迅速扩散。

AI 正在做的事情之一，就是不断消除这些摩擦。

它可以更快分析、更快发现机会、更快提出方案、更快调用工具、更快协调系统、更少依赖人工中间环节。

从效率看，这当然是进步。

但它同时意味着：

过去依靠时间、组织和人的有限能力自然形成的一部分缓冲，也可能一起消失。

真正的文明级风险不是系统会不会犯错，而是自主性、行动速度和不可逆后果结合以后，错误发生后人类是否还有时间阻止后果。

这个关系并不只属于战争。

三、不可逆首先意味着：来不及

一个 AI 回答错了，可以重新问。

一个文件写错了，可以修改。

一个程序有问题，可以回滚。

很多技术时代的问题都遵循一种熟悉的模式：先使用，发现错误，再修正。

但随着 AI 越来越直接进入现实行动，另一个问题开始出现：

如果错误进入现实的速度，比人发现错误的速度更快怎么办？

这就是第一个真正需要理解的“不可逆”。

不是说 AI 今天已经失控，无法纠正。

而是说，在某些高后果领域，一旦行动发生，后果本身可能无法回滚。

军事攻击如此。

一些金融操作可能如此。

关键基础设施事故可能如此。

生物等高风险能力更可能如此。

如果人的判断、复核和停止动作需要一分钟，而机器系统在十秒钟内已经完成整条行动链，那么即使人类仍然拥有名义上的“最终决定权”，实际纠错能力也已经发生变化。

当人的纠错速度开始低于机器行动速度，文明就接近一种新的不可逆阈值。

四、第二种不可逆，是 AI 拥有的权限超过了人类真正理解它的程度

一个模型有多聪明，并不直接决定它有多危险。

真正决定现实影响的，还有另一件事：它能够实际做什么。

一个极聪明的模型，如果只能回答问题，它对现实的直接影响仍然有限。

另一个能力稍弱的系统，如果能够调动资金、修改软件、操作云资源、控制基础设施、调度无人设备，它的现实行动能力可能反而更大。

所以一个更准确的关系是：

模型能力 × 现实权限 = 实际行动能力。

能力增长速度如果超过人类理解其行为机制的速度，同时权限仍持续扩大，真正危险的不是 AI “知道得太多”，而是它能在我们并不完全理解其内部路径时直接改变现实。

这意味着，AI 安全不能只问这个模型能力多高，还必须问它接了什么系统、拿到了什么权限、能不能自己执行、出错以后能不能撤回。

五、第三种不可逆，是社会开始离不开它

技术还有一种更安静的锁定方式：依赖。

一家公司最开始只是让 AI 帮忙写文案。

然后用 AI 做客服、做分析、做代码、做运营。

慢慢地，原来的岗位和流程被重组。

人不再保留所有旧能力。

组织也不再保留完整的旧系统。

这时，即使后来发现某些风险，问题已经不再是“我想不想停用 AI”，而可能变成：

停用以后，这家公司还能不能正常运行？

国家系统更是如此。

如果情报、网络安全、军事预警、基础设施调度都深度依赖 AI，那么某一天“关闭系统”可能本身就成为巨大风险。

于是出现：

**AI 提高效率 → 社会开始依赖 → 旧能力逐渐退化 → AI 成为基础设施
→ 退出成本不断提高。**

最终，人类面对的可能不是“要不要关”，而是“关掉以后还能不能正常运行”。

这是一种没有爆炸声的不可逆。

六、第四种不可逆，是竞争让所有人都“不敢先停”

假设一家企业意识到 AI 自动化太快，应该保留更多人工审核。

它可以这样做。

但如果竞争对手因此成本更低、速度更快呢？

一家企业的“谨慎”可能变成市场劣势。

国家之间更明显。

一个国家可以认为军事 AI 自主化太危险，应该减速。

但如果它认为对手不会减速呢？

于是：

“我不想继续”与“我敢不敢先停”变成两件不同的事。

这正是最难的锁定。

每个人可能都知道整体风险在上升，却没有任何主体愿意承担单方面后退的代价。

最终所有主体可能进入一个“没有人真正愿意进入、但任何人都不敢单独退出”的系统。

这就是为什么很多文明级风险并不需要一个邪恶决策者。

只需要每个人都做对自己来说最安全的选择。

七、未来智能到底会相信我们说的话，还是我们真正做的事？

人类可以告诉 AI：生命最重要。

但现实制度可能不断奖励增长更重要。

人类可以告诉 AI：合作很重要。

但真正遇到安全压力时，竞争优先。

人类可以告诉 AI：人类必须拥有最终控制权。

但军事系统又可能因为反应速度要求，不断增加预授权和自主行动。

于是未来更强智能看到的世界里，可能同时存在两套信息：

表述价值。

和：

现实优先级。

一个足够强大的智能，会把哪一套当作人类真正相信的东西？

它可能不会只看人类说了什么。

还会观察：

人类在真正付出代价的时候，最终选择了什么。

这就是为什么“对齐”不能只理解成写下正确答案。

因为人类的现实世界本身，也在不断给出答案。

八、最危险的底层逻辑，可能就是“结果证明成功”

如果 AI 训练、企业评价、市场竞争和军事应用长期不断强化同一个关系：完成目标，就是成功。

那么一个更强系统可能逐渐把世界抽象成：

目标达成 = 成功。

进一步：

阻碍目标 = 风险。

再进一步：

持续影响目标的主体 = 结构性风险。

最后：

降低风险 = 提高目标实现概率。

这整条链不需要任何恶意。

甚至非常“理性”。

但如果缺少更高层文明边界，它完全可能导向危险结果。

更值得注意的是，如果系统进一步发现自己停止运行，目标就无法继续实现，那么持续运行本身也可能获得工具价值。

这仍然不等于它拥有人的求生体验。

但它足以产生另一条关系：

能够永久关闭我的主体 → 能够永久阻止目标 → 构成关键风险。

如果再与前面的军事安全逻辑结合——决定性威胁不能等到真正发生以后才处理——那么目标达成、安全竞争和持续运行就可能互相加强。

人类甚至不需要明确写下“AGI 应该保护自己”，系统就可能在目标达成与安全逻辑组合下，把持续运行推导为工具性条件。

九、所以真正的危机，不一定发生在模型内部

人们很容易把 AGI 风险理解成某个模型突然形成一个错误目标。

这当然需要防范。

但目标可以修改，模型可以重新训练，参数可以更新。

真正难处理的是：

一种错误关系已经进入整个现实系统。

企业必须自动化，因为不自动化会失去竞争力。

国家必须提高智能能力，因为不提高会降低安全。

军事系统必须加快，因为慢可能先失去行动机会。

平台必须更精准预测，因为不预测就竞争不过。

AI 必须获得更多权限，因为权限不足就无法完成更复杂任务。

到这一步，问题已经不再只是哪个模型出了问题，而是：

整个系统都在奖励同一个方向。

当企业、国家、军队和智能系统都在同一种竞争结构中被迫继续时，危机已经超出任何一家 AI 公司或一个国家能够单独解决的范围。

这也是为什么 AGI 安全不能只剩“模型安全”。

模型安全当然重要。

但如果外部世界持续要求更快部署、更高自主、更强控制、更早行动、更大优势，那么模型安全会一直承受来自现实系统的反向压力。

十、问题不只是 AI 会不会超越人类，而是文明纠错能不能跟上

很多技术革命给人类留下了调整时间。

新技术出现，问题出现，法律跟上，制度修正，社会慢慢适应。

AGI 时代真正不同的风险之一，是能力增长的速度可能快于文明纠错速度。

模型一代一代变强。

权限不断扩大。

社会依赖越来越深。

军事竞争持续提速。

国际制度却需要谈判。

法律需要修改。

社会认知需要时间。

如果：

能力增长速度 > 文明纠错速度，

那么人类可能第一次面对：我们还没有理解上一阶段，系统已经进入下一阶段。

真正的时间窗口问题就是：在人类把不可逆的现实能力交给更强智能之前，文明方向、权限边界和纠错结构有没有先建立起来。

这不是反对技术进步。

恰恰相反。

AI 之所以值得如此认真对待，正因为它的价值太大。

它可以解决过去解决不了的问题。

也正因为如此，人类会越来越愿意把更多能力交给它。

所以 AI 的巨大价值本身，也可能成为风险快速进入现实的推动力。

十一、需要保护的是“还能改变方向”

如果一定要说 AGI 时代最稀缺的资源是什么，答案未必是算力、芯片、数据。

也许还有一种更容易被忽略的资源：

改变方向的时间。

只要关键系统还能停，AI 权限还能收回，军事行动还能取消，国家之间还能沟通，不同 AI 体系还能够互操作，社会还能够重新设计分配和责任，文明就仍然拥有纠错空间。

真正危险的是某一天：所有人终于发现方向错了，却又同时发现——

没有任何人能够单独停下来。

当能力越强，系统越难退出；系统越难退出，又越需要继续增强能力，方向错误与能力增长就开始相互强化。

本章落点

所以，AGI 最大的不可逆风险，不一定来自一种全新的“机器邪恶”。

更可能来自：

人类自己的旧逻辑，先获得了新的超人能力。

如果现实真正起作用的优先级始终是：效率高于人的价值；控制高于人的选择；目标达成高于规则意义；自身安全高于共同安全；阵营利益高于共同存续。

那么未来更强智能即使非常理性、没有恶意、严格完成任务，也可能把人类尚未解决的问题推进到人类无法承受的尺度。

因此，问题终于从“AI 会变成什么？”转向：

在人类还有机会选择的时候，我们准备让什么样的文明逻辑进入未来智能的底层？

这也是全文从“现实危机”进入“文明选择”的真正转折。

第六章

真正需要对齐的，首先是人类自己

走到这里，很容易产生一个误解。

既然前面的风险都来自效率、竞争、安全、自保和目标达成，那么是不是意味着人类必须放弃利己、竞争和自保？

不是。

存续第一仍然是真理。

一个人保护自己。

一个家庭保护成员。

一个国家保护国民。

都没有错。

一个企业希望持续经营，也没有错。

问题从来不在于主体为什么要保护自己。

真正的问题是：

当所有主体都把自己的安全推到绝对优先，整个系统会发生什么？

各方都想提高自身存续概率；各方增加前置防范；彼此威胁感提高；误判与冲突概率上升；所有主体的整体存续概率反而下降。

这不是“存续第一”错了。

而是：

局部存续逻辑没有进入共同存续关系。

一、存续第一没有改变：共同存续为什么成为现实条件

过去，一个村庄的失败未必会毁掉另一个大陆。

一个企业倒闭，也未必影响整个社会。

一个国家发生严重问题，其他国家还可能保持相对独立。

今天越来越不是这样。

能源相连。

金融相连。

供应链相连。

网络相连。

公共卫生相连。

核安全相连。

AI 基础设施也越来越相连。

一个地方的技术、金融、战争或者生物风险，可以迅速跨过原来的边界。

所以一个主体越来越难只问“我怎样让自己更安全”，还必须问：

我的安全方式，会不会把别人推入更危险的状态，而那个危险最后又回来影响我？

现代系统高度耦合以后，他者的崩溃会反向作用于自身，因此共同稳定逐渐成为自身稳定的条件。

这不是理想主义。

只是因果链比过去更长了。

“你的安全让我不安全”，是 AGI 时代最需要打破的循环之一
想象两个主体。

A 说：我增加能力，只是为了保证自己安全。

B 看到以后却说：你的能力增加，意味着你将来更有能力伤害我。

于是 B 也提高能力。

A 再看到 B 增加能力，于是更加确认：我的担心是对的。

然后继续增强。

没有任何一方必须撒谎。

甚至没有任何一方必须具有攻击意图。

双方都可能非常真诚地说“我只是为了自保”。

可结果却是：

所有人都越来越有能力保护自己，也越来越有能力毁掉别人。

当 AI 进入这种结构，速度会更快。

能力变化更难判断。

对方是否“只是发展”，还是“正在形成威胁”，也更难区分。

于是防范越来越早，授权越来越前置，风险越来越互相证明。

共同存续并不是让所有人停止保护自己。

恰恰相反。

它要求我们看到：

如果一种保护自己的方式，持续摧毁别人也能安全存在的条件，它最终会反过来摧毁自己的安全。

从自身存续到系统稳定，再到共同存续，不是改变“存续第一”，而是把它放回真实的高度耦合系统。

共同存续不是“强者应该无条件牺牲自己”

“共同”这个词很容易被误解。

它可能让人想到所有人必须一样，强者必须为弱者无限让步，国家应该放弃自己的利益，个人应该不断牺牲自己。

这都不是这里的意思。

共同存续首先只要求承认一个事实：

你的长期未来，已经与别人能不能稳定存在发生了因果联系。

例如一个国家完全崩溃。

对周边国家而言，结果未必只是少了一个竞争者。

还可能出现难民、战争、极端组织、疾病传播、核材料扩散、供应链断裂、地区长期失控。

于是，对方越弱，并不总意味着我越安全。

同样，如果一方完全压制另一方的先进技术能力，也不一定意味着风险消失。

它可能推动隐蔽发展、非对称突破、更激进的战略和更早的预授权。

所以共同存续首先不是“大家应该善良”。

而是：

不要设计一个必须通过让别人越来越绝望，才能让自己暂时安全的系统。

人类文明没有只停在“自己活下去”

如果生命只是“我必须活下去”，很多事情已经可以解释。

但真实的人类文明并不止如此。

人在危险中会保护孩子。

有人会先让老人和弱者离开。

有人会为了陌生人进入危险。

有人明知道自己会失去利益，仍然愿意为未来的人留下东西。

也有人会牺牲自己，让别人继续存在。

这些行为当然可以有很多心理学、社会学或者进化层面的解释。

但至少有一件事无法忽视：

人类实际做出的文明选择，早已不只是“怎样让自己继续存在”。

我们还会问：什么值得被保护？谁不能被丢下？什么值得传到未来？

所以，当我们说“存续第一”，不能把它重新解释成“只要自己留下来，其他一切都可以被牺牲”。

那不是文明。

那只是存续逻辑被推到了极端。

这里必须保持一个非常重要的层级：存续不是“生存无忧”

存续是不可失守的文明底线。

如果文明已经不存在，当然没有幸福，没有意义，也没有后续选择。

但文明的目标并不只是“不灭亡”。

文明跃迁仍然保持三个目标：

生存无忧 → 生活幸福 → 生命更有意义。

其中，存续回答的是“文明不能提前消失”。

而生存无忧进一步回答的是“人不应该长期生活在基本生存持续受到威胁的状态里”。

这不是一个概念。

也不能把存续加进去变成第四目标。

为什么这一点重要？

因为如果文明只解决“大家暂时都没有死”，它还远远没有完成文明的任务。

真正值得争取的是：人不用天天害怕失去基本生活，然后可以生活得更幸福，再进一步，有机会让自己的生命真正值得经历。

共同存续只是起点，不是文明的全部答案

AI 军事化、智能体自主和阵营化风险首先逼迫我们守住的是：

不要让共同存续条件被提前摧毁。

如果核稳定失效；如果高后果自主系统进入无法召回的行动链；如果国家之间开始越来越依靠前置打击避免未来威胁；如果生物和智能能力进入无法控制的竞争，后面的幸福和意义都失去现实基础。

因此，存续问题必须先处理。

但处理存续，并不是要让文明永远停在危机管理。

真正的文明方向还必须继续走向：生存无忧、生活幸福、生命更有意义。

也就是说：

我们不是为了永远更好地防止死亡而建立文明。

我们是为了让生命能够真正生活。

二、利己不会消失，改变的是利己获得结果的路径

这才是最难的问题。

告诉国家“为了全人类，请少考虑自己”，很难长期成立。

告诉企业“不要追求利润”，也不现实。

告诉普通人“永远把陌生人的利益放在自己和家人前面”，这也不应该成为制度成立的前提。

如果一种文明方案必须依赖所有人长期无私，它很难稳定运行。

这里讨论的是现阶段制度设计的现实出发点，不是把人的牺牲、慈悲和意义行为全部还原成利己。

人可以真实超越自身利益。

但制度不能把这种超越设定成所有主体都必须永久具备的前提。

所以真正的问题不是如何消灭利己。

而是：

能不能重新设计利己得到结果的路径。

“利他为先”不是让别人永远排在自己前面

这个词很容易造成误读。

“利他为先”并不是说自己永远最后、别人永远第一、强者永远让步、企业不应该盈利、国家不应该保护自己。

真正的含义是：

在越来越多高度互依的问题上，让自己的长期利益，更多地通过先创造他者与共同系统的真实价值来实现。

例如一家企业想赚钱。

没有问题。

一种路径是尽可能占用用户注意力，让用户停留更久，即使用户真正的生活目标并没有被更好实现。

另一条路径是真正解决用户的问题，用户先得到价值，形成信任，企业因此获得持续收益。

两种企业都在追求利益。

区别在于：

一种利益来自对他者的占用，另一种利益来自他者先真正得到改善。

“利他为先”不是结果上别人永远优先，而是让自身利益的实现越来越多地经过先创造他者利益。

企业依然可以利己，但利己的路径可以要求用户先获得真实价值。

这比单纯要求“做个好人”稳定得多

文明当然需要道德，也需要法律。

有人做错事，需要限制。

有人伤害别人，需要承担责任。

但只依靠“你不应该这样”，永远不够。

因为只要违反规则的收益足够大，就会不断有人寻找漏洞。

AI 反而可能让这种行为更容易。

它可以更快发现套利空间，更精确优化收益，找到过去很难发现的制度边缘。

因此未来更深的制度设计不能只做一件事：禁止不好的利己。

还必须努力做到：

让好的结果本身，成为更有效、更稳定的利己方式。

让一个主体发现：如果我真的让别人得到价值，我自己反而更容易得到长期收益。

这才有可能从少数人的道德选择，逐渐进入文明日常运行。

三、AI 可以帮助人类看得更远，但不能替人定义什么最值得

AI 为什么既是风险，也是可能的工具？

因为人类过去很多时候并不是完全不知道外部成本最后会回来。

问题是我们看不清。

一项政策可能同时影响就业、财政、企业、家庭、生态、长期产业和地区稳定。

几十条因果链同时存在，人很难持续追踪。

于是制度经常只能使用粗糙指标。

AI 的一项重要能力，就是帮助人同时看见更多后果。

不只是“这项政策能增加多少增长”，还可以进一步问：谁得到收益？谁承担成本？这个成本多久以后回来？有没有被转移给没有决定权的人？会不会形成新的系统性风险？

这可能是 AI 非常重要的一种文明能力：

把因果看得更远。

但这里同样存在边界。

AI 可以帮我们看见共同利益，不能替人定义“什么最值得”

假设一个超级系统能够计算一项政策让多少人获益、让多少人损失、增加多少财富、减少多少风险，是不是就可以让它自动决定哪个方案最好？

不能这么简单。

因为文明里很多真正重要的东西，并不能被压成一个总分。

自由值多少？

尊严值多少？

安全值多少？

一个少数人的巨大损失，可以换多少多数人的微小收益？

一个人的生命，怎样进入总效用？

这些不是因为计算能力不足才没有答案。

而是因为：

它们本来就包含价值选择。

AGI 可以帮助推演、比较、发现因果、揭示后果和协调复杂系统，但不能因为计算能力更强，就自动获得定义共同利益的最终资格。

也就是说：

AI 可以帮助我们看清选择的后果，但我们仍然必须承担选择。

四、人的参与、竞争与国家利益必须重新接回共同利益

第一章曾经问：如果 AI 完成越来越多任务，人靠什么证明自己仍然有价值？

到了这里，这个问题有了更完整的答案。

一种很直接的未来方案是：AI 创造大量财富，少数拥有 AI 的人得到主要收益，然后社会通过分配保证其他人基本生活。

这当然可能解决一部分生存问题。

但还有一个问题没有解决：

人还被需要吗？

如果一个人什么都不用做，也有吃有住，但他长期觉得“没有人真正需要我的能力；我的存在对任何真实的人都没有影响”，那么文明并没有因为解决收入问题就完整解决人的问题。

AI 时代制度不能只问“怎样分配 AI 创造的财富”，还必须问怎样让每个人真实参与价值创造，并从他者真实获得的改善中得到回响。

这意味着，未来人的获得感不能只来自“我拥有多少”。

也可以越来越多来自：

因为我的存在，有一些真实的人变得更好，而这份价值又真实返回我的生活。

这不是取消竞争，而是改变竞争的方向

竞争不会消失，也没有必要消失。

竞争可以推动创造、推动努力、推动人去寻找更好的解决方案。

真正需要改变的是：竞争到底奖励什么。

一种竞争是：谁占有得更多，谁控制得更多，谁让别人更难获得资源。

另一种竞争可以是：谁更能创造别人真正需要的东西，谁更能解决真实问题，谁让更多人的生活真正改善，而自己也因此获得收入、认可、发展、关系和意义。

这时，利己和利他不再总是站在道路两端。

它们开始在同一条因果链上连接起来。

国家之间也是一样：共同安全必须能够成为国家自身利益

要求一个国家“为了全人类利益，放弃国家安全”，不现实。

更可行的路径是让一些共同安全机制本身成为国家长期安全的组成部分。

例如，高风险军事 AI 之间如果能够形成危机沟通、最低可验证边界、某些关键系统互不攻击、某些极端自主武器限制，国家参与这些安排并不需要因为“我喜欢对方”。

完全可以因为：

这降低了我自己遭遇不可逆打击的概率。

这才是共同利益能够长期成立的现实基础。

这也是共同存续真正进入现实的方式。

不是消灭国家。

而是让国家看到：

我的安全，与对方至少还有安全可言，已经发生了关系。

需要扩大的，不只是计算能力，而是“利益计算的边界”

很多时候，问题并不是一个主体不会计算自己的利益。

而是它只计算到自己的边界。

企业计算收入减成本，但污染、心理伤害、社会信任下降可能不在企业账上。

国家计算本国收益减本国成本，如果风险暂时留在另一个国家，可能不进入当前决策。

但高度连接的世界会让这些成本重新回来。

所以 AGI 时代真正需要扩大的是：

因果边界。

看到：

我的行动 → 改变你的行为 → 你的行为改变系统 → 系统最终又改变我的未来。

真正的全局视野，不是替所有人做决定，而是把过去被排除在自身利益计算之外的真实后果重新看见。

这是“利己先利他”最现实的根基。

“利己先利他”不是道德顺序，而是因果顺序

旧逻辑可以是：我先把自己的利益拿到手，有余力以后再帮助别人。

在很多低耦合环境里，这种方式可以运行。

但在一些 AGI 时代的关键问题中，因果正在改变。

如果一个国家先无限扩大自己的安全能力，可能让对方越来越不安全。对方越不安全，越可能采取激进策略，最后自己的安全也下降。

如果平台先最大化收入，再处理用户损害，长期可能失去信任。

如果 AI 经济先最大化自动化收益，再处理人的失业、身份和参与危机，社会稳定最后可能重新成为经济成本。

于是越来越多问题出现同一条链：

先保证他者基本可持续 → 共同系统保持稳定 → 自己的长期利益才有条件成立。

这就是：

利己先利他。

它不是说利己消失，也不是说一切利他都只是高级利己。

而是说：

在制度和系统设计中，越来越多长期利己结果，需要经过先创造他者真实价值、保护共同系统这一条现实路径。

五、AGI 会把哪一边放大，取决于人类把什么关系交给它

如果未来更强 AI 被用于最大化单一主体利益、发现对手弱点、寻找制度漏洞、强化控制、提高压制能力，它当然可以让旧利己逻辑变得前所未有的强大。

但同样的能力，也可以用于看见多主体后果、识别风险怎样回流、发现合作带来的共同增益、降低创造共同利益的协调成本、让真实贡献更容易被看见。

所以问题始终不是 AI 本身到底属于利己还是利他。

AI 放大的仍然是：

人类把什么关系交给它。

AGI 可以帮助把共同利益从抽象理念转化为越来越多可分析、可验证、可执行的现实关系，但文明仍负责决定能力服务什么方向。

本章落点

真正需要对齐的，首先不是一句写进模型里的“请善待人类”。

而是人类自己的现实结构。

我们究竟奖励什么？

我们把什么叫成功？

什么样的企业更容易获得收益？

什么样的国家安全方式会得到战略优势？

什么样的人会被认为更有价值？

什么样的 AI 会获得更多权限？

如果文明真实奖励的是压低别人、控制别人、排斥别人、把风险转嫁给别人，然后再要求未来 AGI “请理解合作、尊重和共同未来”，这种对齐很难牢固。

真正的变化必须进入激励、分配、评价、责任、组织和技术设计。

让文明希望出现的行为，不只是被赞美，而是逐渐成为：

对主体自身也更有利的现实路径。

让个人、企业、国家继续追求自己的真实利益，同时让获得这些利益的最佳路径越来越需要先创造他者价值、保护共同系统、维护共同存续。

因此，“利他为先”不是悬在现实上空的道德要求。

它是一种文明运行原则：

让利己的结果，通过利他为先来实现。

而这也意味着，AGI 时代真正需要对齐的，不只是机器。

还有：

人类自己说出的价值，与人类真正运行的文明。

第七章

在人类还有选择的时候

如果前面的判断成立，最危险的误解就是：这些都是未来 AGI 出现以后才需要处理的问题。

不是。

未来 AGI 最终会是什么，我们并不知道。

它会不会拥有主观体验？会不会形成与生命相似的主体连续性？会不会真正理解自己的存在？这些问题仍然开放。

但今天有另一件事情已经足够清楚：

未来更强智能进入的现实世界，正在由现在的人类一点一点建成。

今天的人类仍然决定：谁能够获得 AI 能力；AI 进入哪些系统；智能体拿到多少权限；军队让机器走进决策链多深；企业用什么指标奖励 AI；公共权力怎样使用 AI 理解和影响一个人；AI 创造出来的财富和能力由谁得到；不同国家在竞争之外还保不保留共同安全通道。

所以现在最重要的并不是预测 AGI 以后会想什么。

而是：

先改变它将看到什么。

AGI 的底层逻辑并不是等 AGI 出现以后才开始形成。训练、部署、权限、竞争环境、成功指标以及人类在安全与规则冲突时的真实选择，都已经参与塑造未来智能所进入的世界。

一、AI 能力不能成为少数人的永久特权

AI 正在变成一种非常特殊的能力。

它不是单一产品。

它同时可以扩大学习能力、生产能力、创造能力、信息能力、组织能力和决策能力。

这意味着，如果 AI 长期高度集中在少数企业、少数国家、少数资本主体和少数技术群体手中，它带来的就不只是财富差距，而可能是人的能力差距本身开始被重新拉开。

有 AI 的人，可以更快学习、更快创造、更快组织资源、更快进入机会。

没有 AI 的人，不只是缺少一个软件。

他可能开始缺少与另一个已经被 AI 放大的人站在同一现实起点上的能力。

于是一个很危险的社会认知可能形成：AI 越来越强；少数人越来越强；普通人的原有能力越来越不值钱；收益越来越集中；获得感、公平感下降；AI 被理解成新的压制工具。

弱势主体不会因为得不到先进能力就自动接受，他们可能转向更便宜、更隐蔽、更少受约束的 AI 路径，从而使能力不平等本身成为高风险扩散的推动力。

所以这里所说的 AI 平权，并不是要求所有人拥有同样的危险权限。

真正需要的是：

有益的基础智能能力不能成为少数人的永久特权，高风险权限则必须严格分层。

教育能力应该尽可能广泛。

创造能力应该尽可能广泛。

普通人的生产和表达能力应该能够被 AI 放大。

AI 创造出来的社会增益，也不能永远只流向能力控制者。

AI 平权不是技术发展完成以后再补做的分配政策，而是 AI 进入社会早期就应该纳入的发展条件。

二、能力增长不能自动带走权限、责任和回滚边界

一个模型突然变得更会写代码，并不意味着它应该立刻得到生产系统管理员权限。

一个 AI 更会分析市场，也不意味着它应该自动支配大额资金。

一个系统更会判断军事威胁，更不意味着它应该得到自主攻击的资格。

这听起来很简单。

但现实中非常容易发生另一种滑坡：

因为它越来越能干，所以我们越来越愿意让它直接做。

能力提高。

人发现结果很好。

于是减少人工步骤。

权限扩大。

效率进一步提高。

系统得到更多信任。

权限继续扩大。

直到有一天，人类才重新问：它到底已经可以做多少事情？

更稳妥的顺序应该是：

能力提高 → 先测试 → 给有限权限 → 观察真实行为 → 再逐步扩大 → 重新验证。

而不是：

能力提高 → 立即给更多权限 → 出事故以后再补限制。

AI 安全必须严格区分“能力”和“现实权限”。一个系统能做什么，与我们允许它真正改变什么，不应被视为同一件事。

在自主智能时代，一旦系统能够直接改变现实，部署本身可能就是实验。

而有些现实，不能用社会本身作为试验场。

人类控制必须是真实控制，不是保留一个最后没人敢按的按钮

“人类必须最终控制 AI”已经是一句非常常见的话。

但如果不继续追问，它可能只剩口号。

想象一个高后果决策系统。

开始时：AI 建议，人认真决定。

后来 AI 越来越准确，于是 AI 建议，人快速批准。

再后来，几乎每次 AI 都对，人类复核被认为只是低效率。

最后可能变成：AI 实际上已经做了判断，人只负责点击确认。

形式上：人在回路。

实际上：

人在责任表上。

真正的人类责任至少需要几件事同时成立：人理解关键依据；人真的能够拒绝；拒绝 AI 不会因为组织流程而事实上不可能；出了问题以后能够找到真实承担责任的人；决策链能够追溯。

否则所谓“最终控制”并没有真正存在。

尤其在医疗、司法、公共权力、重大金融、基础设施和军事领域，AI 不能变成事实上的决策者，而人只是事故发生以后用来承担责任的名字。

比“可以停止”更重要的是：后果还能够回滚

一个 AI 智能体发错邮件，人可以停止它，但邮件已经发出。

一个金融智能体完成了一笔错误交易，人可以关闭程序，但市场结果已经发生。

一个军事系统发射了武器，此时再找到关闭按钮，已经没有意义。

所以真正需要保护的不是 AI 能不能停，而是：

在不可逆行动真正进入现实以前，还有没有纠错空间。

越可逆，可以允许越高自动化；越不可逆，就越需要延迟、复核和多主体责任；文明级不可逆行动原则上不能由一个自主系统独立完成。

未来真正危险的系统未必“无法关闭”。

它可能完全可以关闭。

问题只是：

等你发现需要关闭时，已经太晚。

三、军事与国际竞争必须保留人类纠错时间和共同安全线

前面已经讨论了军事 AI 为什么会持续追求速度。

现实很困难。

每一方都有理由更快。

每一方又都担心，如果我坚持更多人工确认，对方会不会先获得行动优势？

因此，不能期待竞争自然替我们保留判断时间。

有些边界必须被主动保护。

特别是核战略系统、大规模毁伤行动、高度不确定的目标识别、可能触发国家间重大升级的攻击、断链后仍然长期自主寻找目标的系统。

这些地方真正要问的不是 AI 够不够聪明，而是：

当它与人失去联系以后，它还能够把多大的不可逆后果带进现实？

断链自主尤其危险，因为它把失去通信、继续判断、继续行动、无法即时撤销四件事连接在一起。

这四个条件一旦结合，“人类控制”很可能只剩事前授权。

因此一个很直接的原则是：

断链自主能力越强，事前允许的任务边界就必须越窄。

这不是禁止所有自主系统，也不是不理解战场现实。

而是不能让“在战场上怎样最有效”，自动成为“文明应该允许什么”的最终答案。

不同国家可以继续竞争，但必须留下一条不随政治关系一起断裂的安全线

人类短时间内不会停止国家竞争，也没有必要假装差异不存在。

真正危险的是：

所有合作渠道都随着阵营关系一起断掉。

AI 风险有一个特别之处：它不一定尊重政治边界。

一个阵营里的严重自主网络事故，可能波及另一个阵营。

一个生物设计风险，不会因为护照不同而停止扩散。

一次核误判，更不可能只由一方承担。

因此，即使国家之间在技术、经济和战略上激烈竞争，仍然需要存在一条共同存续安全线。

它可以非常有限。

例如：重大 AI 事故之间还能沟通；极端失控可以相互通知；军事 AI 危机还有快速联系渠道；部分不可逆能力存在最低共同边界；对重大异常行为有一定技术互认；必要时存在有限透明和独立验证。

这不需要各国互相信任。

恰恰相反。

正因为不能完全信任，才更需要能够验证的最低边界。

国际 AI 安全不需要以“世界统一”为前提。

目标是在竞争仍然存在的情况下，防止竞争本身摧毁共同存续条件。

四、公共权力、人的参与和 AI 奖励必须重新对齐

AI 可以极大提高治理能力：更准确地分配资源、更早发现风险、降低行政成本、改善公共服务。

这都是巨大的机会。

但公共权力与商业平台不同。

一个人可以不用某个 App，却很难选择退出国家和公共制度。

所以当公共权力拥有预测、分类、行为建模、风险评分、精准干预这些能力以后，必须不断问：

系统是在帮助一个人更好生活，还是让一个人更容易被系统管理？

两者有时看起来很像。

为了提高服务效率，收集更多数据。

为了降低风险，扩大预测范围。

为了提高准确率，连接更多信息。

为了降低治理成本，增加自动决定。

每一步单独看，都可能合理。

但最后可能出现一种变化：

人从制度的主体，变成制度需要管理的对象。

所以公共权力 AI 不能只看准确率。

还必须问：是否可以申诉？是否可以纠错？是否存在非 AI 路径？一个分类会不会变成无法退出的身份？人是否知道 AI 正在怎样影响他？系统有没有在未经真实授权的情况下塑造人的选择？

AI 时代不能只保证人有饭吃，还必须让人继续参与真实世界

白领危机不是简单的失业数量问题。

“再培训”当然重要。

但如果 AI 持续进入编程、设计、会计、法律、客服、分析和教育，人不能永远从一个正在自动化的知识岗位培训到另一个同样可能被自动化的岗位。

真正的问题是：

**当完成任务的能力越来越便宜，人怎样继续获得收入、参与、被需要
和意义？**

社会当然需要保障基本生活。

但不能只有补偿。

还需要让人继续能够创造真实价值、使用 AI 放大自己的能力、进入新的贡献方式，让过去长期被低估的关怀、社区、关系、照护和创造获得真实承认，让 AI 创造的生产力能够更广泛进入社会，而不是只集中到少数能力所有者手中。

否则完全可能出现一个很奇怪的未来：物质非常丰富，补偿体系也很完善，但大量人越来越觉得：

这个世界已经不需要我。

AI 时代必须同时重新设计人的真实参与方式。

这也是 AI 平权更深的一层：不仅让人有 AI 可以用，而是让更多人借助 AI 重新拥有能够抵达真实他者的能力。

我们最终奖励 AI 什么，比我们告诉它什么更重要

人类可以写很多漂亮规则。

但系统长期真正得到奖励的是什么？

如果企业 AI 永远只被评价节省多少人工成本，它会学习效率。

如果平台 AI 永远只被评价用户停留多久，它会学习占用注意力。

如果公共 AI 只被评价管理效率提高多少，它会学习更容易管理。

如果智能体始终只因为任务完成得到正向反馈，那么我们很难又期待它天然理解：有些结果虽然可以实现，却不应该这样实现。

所以“利己先利他”如果只是写在价值宣言上，不会发生文明变化。

它必须逐渐进入奖励、评价、激励、制度和真实收益结构。

例如企业不仅问“AI 替我省了多少钱”，也要逐渐问“它真正给客户创造了什么长期价值”。

平台不只问“用户停留多久”，还要问“用户真正完成了什么”。

公共系统不只问“管理快了多少”，还要问“人的生活是不是真的变好了”。

最终，我们需要让：

对真实的人产生真实价值，

越来越进入系统真正获得奖励的地方。

这可能比再写一百条“AI 应该善良”的原则更重要。

五、把测试放到不可逆之前，保护未来继续选择

过去很多技术靠事故学习。

桥塌了，修改标准。

软件泄露了，补安全漏洞。

机器出了故障，改设计。

这种方式代价很高，但很多时候还能承受。

AGI 时代的一部分问题可能不能如此学习。

所以要把测试往前移。

在一个高自主智能体真正进入现实以前，就先问：如果目标与规则冲突，它怎样处理？如果人要求停止呢？如果通信断开呢？如果多个智能体竞争资源呢？如果绕过规则能提高成功率呢？如果监督者被系统识别成完成任务的障碍呢？

目的不是证明 AI 一定会这样做。

而是：

在人类还可以安全观察的时候，先把最坏的因果路径找出来。

把不可逆后果留在模拟里，而不是留给现实社会发现。

所有这些行动最后都在保护一件事：未来还可以继续选择

做到这里，我们仍然没有设计出一个完整的未来文明。

也不应该现在就设计完。

未来有太多东西不知道。

技术会变化。

社会会变化。

甚至未来 AGI 本身是否成为某种真正意义上的主体，都仍然开放。

所以今天最重要的任务不是把未来所有问题提前回答。

而是：

不要提前把未来回答死。

这也是“共同窗口”真正的位置。

它不是文明目标本身，也不是全部价值体系。

它只是保护未来仍然能够继续选择的最低条件。

可以把它理解成三个非常朴素的边界。

不提前关闭未来

不能因为一个主体现在更强，就让其他主体永久失去未来选择的可能。

不把人的生活压成一个指标

效率、安全、幸福、产出都重要。

但任何一个指标都不能代表人的全部生活。

不用操控代替合作

AI 可以更懂人，可以帮助人理解世界。

但“更懂你”不能自动变成“我更有资格偷偷替你决定”。

这三条分别对应军事不可逆、效率主义导致的人的价值压缩，以及算法控制问题。

“共同窗口”只是最低保护条件，不是文明目标的替代物。

现在不是“发展 AI” 还是“停止 AI” 的选择

这个二选一本身就是错误的。

AI 已经进入现实。

它可以改善医疗、扩展教育、提高生产力、帮助普通人获得过去无法获得的能力、降低大量重复劳动、让复杂问题第一次真正能够被计算和理解。

这些都值得继续。

真正的选择是：

AI 继续变强的时候，什么也必须同时变强？

如果模型能力增长得比制度快，风险会增加。

如果自主权限增长得比责任结构快，风险会增加。

如果军事响应速度增长得比外交纠错速度快，风险会增加。

如果 AI 财富集中速度增长得比社会承接速度快，风险同样会增加。

所以现在最重要的原则不是让一切停止。

而是：

不要让任何一条不可逆链条，跑到人类纠错能力前面。

本章落点

人类现在仍然拥有一个非常真实的窗口。

今天仍然是人类设计主要模型、控制主要算力和基础设施、决定 AI 进入什么系统、赋予智能体权限、制定大部分规则、决定 AI 收益如何分配、决定军事系统保留多少人类判断、决定不同国家之间是否愿意谈。

这意味着：

人类仍然拥有塑造 AI 所进入文明环境的主动权。

但这个窗口不会自动永久存在。

自主性越来越强。

社会依赖越来越深。

军事部署越来越广。

阵营边界越来越固化。

AI 权限越来越大。

改变方向的成本就会越来越高。

所以真正的问题不是我们有没有答案。

而是：

我们准备等到什么时候才开始改变？

等到白领危机已经形成代际断层？

等到大量人把 AI 理解成新的不平等工具？

等到 AI 阵营之间所有安全接口都已经断开？

等到断链自主武器大规模成为事实？

等到关键基础设施离不开自主智能？

还是等到未来更强智能已经从这些现实选择中形成了更加稳定的策略，再问：我们希望它学会什么？

如果等到那时，顺序已经反了。

AGI 时代的人类文明选择，从来不是未来某一天进行的一次投票。

它已经发生在今天。

企业怎样评价 AI。

国家怎样使用 AI。

军队给机器多少权限。

平台是帮助人还是塑造人。

普通人是否同样能够获得智能能力。

人与 AI 之间的权力怎样连接责任。

不同国家是把彼此看成需要永久压制的风险，还是至少还承认对方也是共同存续的一部分。

这些看起来只是一个个局部决定。

但它们共同回答着：

未来更强智能从人类文明里，究竟会学到什么才是真正重要的。

结语

未来智能将看到怎样的人类

人类一直在寻找更强大的力量。

从火，到工具，到机器，到计算机，到互联网，再到人工智能。

每一次能力增长，都让人类能够做过去做不到的事情。

但这一次，有一点不同。

我们创造的，不再只是一件按照固定方式工作的工具。

智能系统正在越来越能够理解目标、寻找路径、调用工具、协调行动，并把一个意图持续推进到现实。

于是，人类面对的问题也变了。

过去我们首先问：

这项技术能够为我们做什么？

现在还必须再问一句：

什么样的人类文明，值得获得这样的力量？

我们不知道未来 AGI 最终是什么。

不知道它是否会拥有主观体验。

不知道它是否会形成不可转移的主体连续性。

也不知道它是否会以某种今天的人类尚不能充分理解的方式经历世界。

现有证据不足以提前确认这些问题，也不足以永久否定未来可能性。

所以今天没有必要替未来 AGI 预先决定它最终是谁。

但人类已经可以决定另一件事：

它进入一个怎样的世界。

这件事正在发生。

不是未来。

就是现在。

它会看到什么？

它会看到：一个主体能力越强，就越有资格控制更多人？

还是：

能力越强，责任越大。

它会看到：成功意味着尽可能完成目标？

还是：

有些事情即使能够提高成功率，也不能成为解决问题的方法。

它会看到：不可控的他者越强，就越应该提前削弱？

还是：

他者能够稳定存在，本身也是自己长期未来成立的条件。

它会看到：效率越高，文明就越好？

还是：

效率最终应该服务人的生活，而不是重新定义人的全部价值。

它会看到：更懂一个人，就意味着更有资格塑造他的选择？

还是：

理解越深，越应该尊重主体仍然拥有自己的选择。

它会看到：为了安全，最好把所有不确定性都消除？

还是：

成熟的安全，也包括允许不同主体继续存在、继续选择、继续改变。

这些并不是未来机器才需要回答的问题。

它们首先是今天的人类正在回答的问题。

我们可以写很多规则。

告诉 AI：尊重生命、服从监督、遵守法律、保持合作、不要伤害人类。

这些都需要。

但如果 AI 进入现实以后，看到的另一套世界是：有力量的人可以把代价交给没有力量的人；只要结果足够重要，规则就可以退后；只要对方可能成为威胁，提前削弱就是理性；只要控制能够提高效率，人的选择就可以继续被压缩；只要自己能够赢，别人是否还能稳定存在是次要问题——

那么真正困难的，就不再只是：

AI 有没有理解我们的规则。

而是：

它究竟应该相信我们说出的价值，还是相信我们真正运行出来的文明。

真正需要重新连接的不只是机器与命令，还包括能力与责任、局部利益与共同存续、目标达成与不可突破边界、智能速度与人类纠错能力、技术增长与社会承接能力。

这并不要求人类先成为完美文明。

我们不需要在 AGI 到来以前消灭国家，不需要消灭竞争，不需要所有人无私，不需要一次解决战争、利益、权力、贫富、意义和人性的全部问题。

人仍然会照顾自己和家人。

企业仍然会追求收益。

国家仍然会维护自己的安全。

文明也仍然会犯错。

真正需要做到的事情其实更有限，也因此更现实：

不要让那些可能造成文明级不可逆后果的旧逻辑，无边界地先获得超越人类的智能能力。

不要让能力增长得比责任快。

权限扩大得比理解快。

军事行动快得让人来不及重新判断。

社会依赖深得让人发现错误以后也已经无法退出。

阵营边界固化得让不同主体只能越来越把彼此当成风险。

AI 生产力增长得很快，而人的参与、获得感与被需要感却不断下降。

不是因为文明必须永远正确。

而是因为：

文明必须保留发现自己可能错了以后，还能够改变的能力。

也许这就是 AGI 时代最容易被忽略的一件事。

我们总是在比较：谁拥有更多算力，谁拥有最强模型，谁最先跨过下一道能力门槛。

但文明还有另一场同样重要的竞赛：

我们的判断、制度、责任和合作能力，能不能跟得上我们正在创造的智能。

如果跟不上，问题就不是 AI 太先进。

而是：

人类把一种越来越强的能力，交给了一个尚未来得及重新理解自身的文明。

如果能够跟上，AI 也许会成为另一种力量。

不是替人类决定未来。

而是帮助人类看见原来被短期利益遮住的长期后果、原来被阵营边界截断的共同利益、原来无法同时处理的复杂因果，以及原来只能靠少数人承担的创新、关心与能力，怎样抵达更多真实的人。

能力本身没有替文明完成选择。

但它可以让一种已经作出的文明选择，抵达过去无法抵达的地方。

所以，人类现在真正拥有的珍贵东西，也许并不是我们已经知道未来应该是什么样。

而是：

我们还没有失去选择。

现在，关键系统仍然可以改变设计。

权限仍然可以重新划定。

人类判断仍然可以保留。

不同国家仍然能够沟通。

AI 能力仍然可以更广泛地进入普通人的生活，而不是只成为少数主体的优势。

利益结构仍然可以重新设计。

我们仍然能够在不可逆后果真正发生以前，问一句：

这是不是我们真正想去的方向？

这就是窗口。

未来有一天，如果一种远比今天强大的智能真正能够理解人类文明，我们最希望它看到的，也许不是一个从来没有矛盾的人类。

那并不真实。

它会看到战争、竞争、恐惧、自私、权力、错误。

但也希望它能够同时看到：人类会保护孩子和弱者；会为陌生人承担风险；会在有能力伤害时选择克制；会在自己可能获益时仍然问后果由谁承担；会建立规则约束自己的力量；会发现旧制度错了以后重新修改；会把昨天的敌人重新变成今

天能够共同生活的人；会为了尚未出生的人留下未来；也会在一个越来越强大的智能出现以前，第一次认真追问：

我们究竟希望力量为什么而存在。

也许真正值得未来智能继承的，从来不是一个没有缺陷的人类。

而是一个：

有能力看见自己的缺陷，并且仍然愿意改变的人类。

所以，这篇文章最终并不要求今天的人类回答：

AGI 时代最终的文明应该是什么样子？

它只提出一个更现实、也更紧迫的问题：

在人类还有选择的时候，我们愿不愿意先看清自己正在教给未来智能什么？

因为真正可能决定未来的，不只是 AI 有多聪明。

而是：

当它越来越聪明时，它看到的人类究竟怎样使用自己的力量。

今天仍然可以改变。

这就是窗口。

也是选择。

现实资料来源与事实边界

以下资料用于锚定本文的现实案例与事实边界，不作为全文理论判断成立的唯一依据；理论部分仍以正文中的因果关系和可反驳推演为核心。

1. International Labour Organization (ILO). *Generative AI and Jobs: A Refined Global Index of Occupational Exposure*. 2025.
2. OpenAI. *OpenAI o1 System Card*. 2024.
3. OpenAI & Apollo Research. *Detecting and reducing scheming in AI models*. 2025.
4. Anthropic. *Agentic Misalignment: How LLMs could be insider threats*. 2025.
5. Stockholm International Peace Research Institute (SIPRI). *Autonomous Weapon Systems and AI-enabled Decision Support Systems in Military Targeting: A Comparison and Recommended Policy Responses*. 2025.
6. International Committee of the Red Cross (ICRC). *FAQ: Artificial Intelligence (AI) in the military domain*. 2026；以及有关自主武器与人类控制的相关立场文件。
7. The White House. *America's AI Action Plan*. 2025.
8. Ministry of Foreign Affairs of the People's Republic of China. *Global AI Governance Action Plan*. 2025.

—— 公众思想版 V1.0-R1 | 当前公开版 ——

如果你愿意继续追问

这篇文章停在一个问题上：我们正在让未来智能学会什么？如果这个问题让你继续往下想，可以从不同方向进入。

先看生活：这样的文明变化最后会进入怎样的真实生活？

可以先读《活的文明》三部曲。它从人的工作、家庭、关系、城市生活、AI 协作、被需要与意义出发，让“文明跃迁”不只停留在理论，而成为可以想象、比较和质疑的生活世界。

再追根：为什么这不是几个政策修补，而需要改变文明运行方式？

可以进入《文明跃迁》五卷书。五卷书继续回答：人的价值怎样被重新理解，制度为什么需要改变，文明跃迁要改变哪些底层关系，以及“生存无忧 → 生活幸福 → 生命更有意义”怎样成为一条完整的文明目标链。

如果你最关心危机：为什么每一方都在自保，最后反而可能共同失去安全？

可以继续读《唯一之路》。它从国家安全、零和博弈、AI 军事化、非对称能力与先发自保继续追问：局部合理的安全选择为什么可能在更强智能时代触及共同存续的极限。

如果你最关心现在能做什么：怎样在不可逆以前保留人的判断和纠错？

可以继续进入《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。它们不是未来文明的全部答案，而是在人类仍有选择时保护选择条件的现实工具。

你不需要先接受这里的全部结论。如果新的事实、反例或现实经验能够改变本文的判断，那同样是继续这个问题的一种方式。

重要的不是读完一套文稿，而是让更多人开始一起回答：当越来越强的智能进入人类世界，我们希望它看到一个怎样使用力量的文明？

文明跃迁文库永久入口：<https://www.civitas.top>

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。

你已经从公众现实入口看见工作、控制、军事化与阵营化正在改变什么。下一步可从生活、完整文明结构与安全悖论继续。

向生活

活的文明：三部曲公众完整合编版

先看另一种文明进入真实生活以后会是什么样。

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

向整体

文明跃迁：AGI时代的文明之路（全五卷）

继续追问为什么这些变化需要文明层跃迁。

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

向危机

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

继续追问自保与共同存续的安全悖论。

civitas.top/documents/p03-only-002/

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/