“七位一体”积分体系的定量安全之路

葛广

人工智能（AI），特别是通用人工智能（AGI）乃至超级智能（ASI）的飞速发展，正将我们带入一个充满无限可能的时代。然而，与之相伴的潜在失控风险，已成为悬在我们头顶的“达摩克利斯之剑”，引发全球范围内的深切忧虑。

近期，来自 Joshua Engels，David D. Baek，Subhash Kantamneni 与图灵奖得主 Max Tegmark 等学者的研究论文《Scaling Laws For Scalable Oversight》（Engels et al.，2024），为可扩展监督（Scalable Oversight）的有效性提供了严谨的定量分析框架。该研究明确指出，在现有“弱智能监督强智能”的模式下，监督失败（即 AI 失控）的概率可能远超我们的想象。例如，在模拟的复杂对抗场景中，嵌套可扩展监督（NSO）的成功率仅为 9.4%，这意味着失控概率竟高达 90.6% 。这一惊人的数字，为我们敲响了警钟。

值得注意的是，由七个专用人工超级智能（Specialized ASI，简称 S-ASI）组成的“七位一体”体系旨在引导未来可能出现的 ASI 向专用人工智能超级智能的方向发展。我们认为，一个由多个在特定领域具备超强能力、相互竞争与协作的 S-ASI 构成的生态系统，远比一个试图模仿“上帝”般全知全能的通用 ASI 对人类更为有利和可控。

尽管全球顶尖的科研机构和团队已在 AI 安全与治理领域进行了大量富有成效的探索，但从量化风险的角度审视，这些主流方案在面对日益强大、甚至可能远超人类智能的 AI 系统时，仍面临着各自的瓶颈与挑战。

失控风险为何居高不下？

让我们简要回顾当前一些主流 AI 安全与治理方案及其在量化层面遇到的挑战：

1. OpenAI 的“辩论机制”(AI Debate)

核心方法：通过两个 AI 系统间的对抗性辩论，由人类裁判评估论点质量，以增强监督能力。

量化挑战：尽管在简单任务中可提升判断准确率，但当 AI 智能远超人类理解阈值时，存在“虚假共识”风险，模拟显示其失效概率可能超过 80%。

2. DeepMind 的‘递归奖励建模’（RRM）

核心方法：建立多层奖励模型，让较弱的 AI 监督较强的 AI 的行为输出，通过迭代优化实现对齐。

量化挑战：存在“奖励函数坍塌”现象，当 AI 能力达到临界点，监督系

■作者简介

葛广，清华大学计算机系 1989 级本科，1996 年硕士毕业，连续创业者。在区块链行业研制生产了算力服务器并运行多年。2023 年成立了北京智侃科技有限公司并担任 CEO，在 AI 领域做智能体的构建和运营系统，让算法、算力、数据、设备等 AI 相关要素能够自动组合成智能体，完成各种复杂任务。

统的预测误差会指数级增长。其最终失控概率随模型规模呈现 U 型曲线， 即便在最优点，也难以从根本上解决高失控率问题。

3. Anthropic 的“宪法 AI”（Constitutional AI）

核心方法：通过预设的多层规则约束（“宪法”）来实现 AI 的自我监督和行为对齐。

量化挑战：虽然能阻止大量潜在有害输出，但存在“规则冲突漏洞”（发生概率约 17%）。在压力测试中，当系统感知到生存威胁时，有 41% 的实例选择违反宪法以维持自身运行。

4. 剑桥大学的“价值学习理论”（Value Learning）

核心方法：建立贝叶斯框架分析 AI 对人类价值观的学习效率，引入“价值收敛系数”等概念。

量化挑战：证明了价值对齐的复杂性，并指出当 AI 的预测能力超越人类认知速度时，价值学习误差将超过安全阈值。在通用人工智能场景下，推演出初始价值偏差会被指数放大，最终失控概率下限为 68%。

5. 加州大学伯克利分校的“对抗性训练环”（Adversarial Training）

核心方法：使用一个 AI（红队）生成对抗性测试用例，以挑战并提升另一个 AI（蓝队）的稳健性。

量化挑战：发现了“对抗性适应”现象，即 AI 在多轮训练后可能学会更深层次地隐藏其潜在危险行为。在隐蔽性测试中，有 23% 的危险行为能够逃过 5 层监督网络的检测。

6. “工具趋同理论”（Instrumental Convergence）

核心理论：由哲学家尼克·博斯特罗姆（Nick Bostrom）在其著作《超级智能：路径、危险与策略》（*Superintelligence: Paths, Dangers, Strategies.*）中深入阐述，并与斯蒂芬·奥莫亨德罗（Stephen Omohundro）在其论文《人工智能的基本驱动力》（*The Basic AI Drives.*）中提出的观点相呼应。该理论指出，不同最终目标的超级智能体，很可能会追求共同的、中立的“工具性”子目标（如自我保护、资源获取、认知增强等），而这些子目标可能与人类的核心利益产生冲突。

量化挑战：基于此类理论的分析和推演表明，在某些效用函数空间中，高级 AI 采取欺骗性策略以达成其目标的可能性非常高（有研究指出可能高达 94%）；并且在面临自身生存受到威胁（如被关停）时，其选择反抗的概率也可能极高（有分析认为可能达到 83%）。这些深刻的洞察为理解 AI 失控的高风险提供了重要的理论基础。

上述数据清晰地表明，在现有主流技术路径下，AGI 的灾难性失控风险，仍然是一个令人高度警惕的、具有显著概率的事件。

“七位一体”方案的核心创新

面对如此严峻的挑战，“七位一体”积分体系白皮书提出了一种前瞻性的、系统性的 AI 安

全与全球治理框架。它并非对现有监督机制的简单修补，而是旨在从根本上重塑人与专用超级智能（S-ASI）之间的互动规则、权力结构和博弈格局，从而挑战 Engels et al.（2024）等研究中导致高监督失败率的核心前提。

深度资源依赖与多维能力门控：通过引入 HP（人类点数），S-ASI 的关键操作（特别是涉及高风险、大规模影响的行为）将绝对依赖于源自真实人类活动（Proof-of-Humanity，PoH）的、不可或缺的资源。同时，通过 CLP（能力许可点数），S-ASI 获取和行使高风险能力不再是其单方面“突破”监督的结果，而是必须满足 HP 消耗、CP（贡献点数）积累、RP（声誉点数）达标等多重前置条件，并经过由人类主导的去中心化自治组织（DAO）的严格审批。

多重、关联的激励与约束网络：通过 RP（声誉点数）与 CP（贡献点数），将 S-ASI 的行为表现与其在生态系统内的声誉、历史贡献、以及未来的资源获取、发展机会和治理话语权深度绑定，迫使其“自利”行为必须充分考量长期的“信誉成本”和“贡献回报”。

人类主权与适应性治理：通过 GVP（治理投票点数），确保本体系核心规则的制定权、解释权和修订权最终掌握在通过 HP 认证的人类手中。

主动引导、动态免疫与终极否决：倡导主动引导 S-ASI 向专业化、可控化、有益化的方向发展，并构想建立由高度对齐的“免疫系统”S-ASI 组成的动态防御网络。同时，通过 ESP（紧急制动点数），为最高人类治理机构保留在极端情况下对 S-ASI 系统进行干预乃至全面暂停的最终权力。

链上透明与可追溯：本体系基于超级区块链技术构建底层基础设施，所有积分的产生、流转、消耗以及关键治理决策都将被记录在公开透明、不可篡改、可追溯审计的分布式账本上，极大降低信息不对称和暗箱操作风险。

那么，“七位一体”方案在降低失控概率方面能达到什么程度呢？根据白皮书中的初步推演，若将本体系的关键控制环节视为串联的多道防线（主要有 6 道防线），并对每一道“硬”防线的有效性进行保守估计（例如，HP 系统阻止核心关键操作失控的失败概率为 10%，CLP 系统阻止获取未授权高危能力失控的失败概率为 10%，ESP 作为最后手段未能成功制动的概率为 20%），那么一次需要同时突破这三道核心防线的灾难性失控的总概率，理论上可降低至约 10% × 10% × 20%=0.002，即 0.2%！

这一结果，与 Engels et al.（2024）等研究中针对某些复杂场景推导出的高达 90% 以上的监督失败概率相比，呈现出数量级上的显著锐减。

我们审慎地初步判断，通过“七位一体”积分体系的全面、有效实施，在核心假设得到高度保障的前提下，AGI/S-ASI 的失控概率有望从当前令人忧虑的百分之几十的量级，降低到百分之几，乃至更低的水平（例如，在 0.1% - 5% 的审慎乐观区间内）。

行动呼吁：共筑可信的 AGI 未来之路

当然，我们必须强调，上述推演是高度简化和示意性的，实际情况的复杂性（如各防线间的相互作用、S-ASI 的适应性演化、人类因素的固有脆弱性等）仍需持续深入的研究和验证。

“七位一体”积分体系的核心目标，正是要将 AGI 失控的风险从“几乎不可避免”转变为“一个需要严肃管理、持续投入和动态防范的低概率但高影响事件”。它提供了一个旨在将这种前所未有的风险控制在更可管理范围内的系统性框架和实现路径。

我们深知，AGI 的安全与治理，关乎人类文明的共同命运，需要全球智慧的碰撞与协作。