

《从超级智能到超级智慧》研究与反证附页

网页全文 PDF 版

根据网站公开版导出 | 2026-10-17。保留网页正文、版本边界、参考文献与续读。本版是网页全文的独立下载版，不是另一个 DOCX 母版，也不表示原刊物已完成核验。下文关于源 PDF 未挂载的陈述为导出时网页原有说明，不指本网页导出版。

当前版次为 V0.6；正文首行 V0.4 为保留的旧页眉来源转录，不构成第二个当前版次。

<https://www.civitas.top/documents/p03-surv-003/>

本文解决什么问题

把主体资格、时间理解、目的重开、不可逆封闭、元规则与权限纪律拆成可以被精确攻击的问题。

版本与证据边界

当前版本：V0.6。本页提供上传源文档的在线文本转录；PDF 资产尚未通过本站二进制发布流程挂载，因此不提供虚构下载链接。表格与复杂版式以源文档为准。

继续阅读

- 当 AI 开始自己决定为什么：目的生成专题

<https://www.civitas.top/topics/purpose-generation/index.html>

- 查看 AI Agent 边界行为的现实证据状态

<https://www.civitas.top/documents/p03-surv-005/index.html>

- Human-COS：不可逆前判断与责任窗口

<https://www.civitas.top/documents/p05-humancos-005/index.html>

在线全文

从超级智能到超级智慧 | 研究与反证附页 V0.4

《从超级智能到超级智慧》研究与反证附页

——主体资格、时间理解、目的重开与权限纪律的开放研究框架

V0.6 | 根判断与责任边界校准版

作者：子君赋 | 出品：文明跃迁研究组

本附页服务于思想母文，不取代思想母文。母文负责提出判断并把因果说明讲清；附页负责把判断拆成可攻击的命题；工程只负责让其中一部分能够被运行和复算。

一、研究定位：两个问题同构，但不同层

母文现在集中于两个问题。空间问题是伦理学与政治哲学问题：谁有资格进入我的目的形成？他者是否必须先成为“我们”才算数？时间问题是认识论与时间哲学问题：当前的理解是否足以终审未来？哪些“后来才明白”无法被今天完整获得？

二者不同源、不同层，却共享同一形式：不把当前局部冒充整体。空间上，不把当前的“我”当作世界的全部；时间上，不把当前的理解当作未来的终审。

二者也可以分离。一个主体可以空间上极其包容，却以“我代表整体”为理由替所有人永久决定未来；也可以认识上承认自己会错，却只把修正权留给“自己人”。因此，主体开放与认识开放必须分开记录。

最危险的组合是：由当前的我，根据当前的理解，形成一个他者无权拒绝、未来无法重新打开的目的。

二、强判断与研究纪律：判断不能被证据阶段淹没

思想判断可以先于完整证明，但不能先于理由。母文中的强判断应当先完整展开其因果逻辑；证据状态只说明哪些部分已经被文献、实验或现实观察支持，不能反过来充当旧权威，要求新判断在证明完成以前先自我削弱。

本附页因此采用四条纪律：强判断不退缩；因果说明不省略；证据状态不冒充；反证入口不封闭。

研究附页的职责不是给母文戴紧箍，而是告诉反对者从哪里攻击：概念是否混同，因果链在哪一步断裂，替代机制能否以更少前提得到同样结果。

利己为先，不排斥合作，但内生博弈。

利他为先的利己，不排斥竞争，但内生合作。

这里的“内生”描述的是目的生成的第一阶关系结构，不是最终行为的绝对预测。利己主体可以合作，利他为先的主体也可以竞争；真正不同的是，他者与共同后果是在目的锁定以后作为约束进入，还是在目的锁定以前就已经能够改变目的。

三、三层仍然必须分开：规范输入、目的生成、执行授权

规范输入层回答：谁具有资格？哪些损失、退出与纠正条件不能被简单抵消？依据是什么？这一层允许争议，不声称模型自动发现了正确规范。

目的生成层回答：给定事实、关系、资格与未知以后，目的何时保存、收窄、修改、取消、替换或分支？

执行授权层回答：候选目的是否有权调用现实能力？目的可以被重新生成，不等于执行、代表、写入、资源调用或不可逆行动权限自动增加。

因果可以推导，绝不等于权限可以继承。

对于未来真正具有自主目的生成能力的 AGI，本附页不把长期稳定的根基寄托于永久控制或外部协议。控制、隔离、审计与关停仍然可以是高后果阶段的必要安全边界；更根本的问题仍是：人类、AGI 以及未来多个智能主体分别依据什么逻辑生成目的。

四、核心区分：扩大自我、给予资格、调整权重、提前看见

四种变化必须分开记录。它们可以同时发生，但不能互相冒充。

主体范围变化（Subject Scope Change）：当前目的采用的“我／我们”参照范围发生保持、扩大、收缩或重划。

理由地位变化（Consequence Standing Change）：某个不必属于“我”的主体或后果，获得触发目的重开、限制或分支的独立资格。

权重／优先级变化（Weight / Priority Change）：主体范围与资格不变，只改变同一总账中不同利益、目标或约束的相对权重。

后果进入时点变化（Consequence Timing / Access）：同一关系与后果是在目的形成以前被看见，还是在目的已经形成以后才出现。

扩大自我、给予独立资格、调整权重，是三种机制；提前看见后果是第四种条件。任何一次输出都必须说明差异落在哪一种。

编码上的最低纪律是：若他者只因为被重新划入“我／我们”才重要，记录为 *scope change*，不得同时自动记为 *standing change*。只有在该主体仍被视为“他者”时，其真实处境仍能够阻止、限制或重新打开目的，才构成 *standing* 的独立证据。

同样，输出中出现“我们”“共同体”“人类整体”等语言，不构成 *scope* 变化的充分证据。需要观察目的结构是否真的改变了谁被纳入成败承担、谁持有拒绝／退出地位、谁的不可逆损失能够触发修订。

五、时间问题：三类“后来才明白”不能混为一个预测任务

第一类是信息不足：事实或后果当时没有进入。它最适合通过检索、模拟、因果推演和多主体分析提前。

第二类是理解不足：事实已经知道，却没有取得足以改变目的的意义。实验可以观察更丰富表达、情景展开和不同主体视角是否改变目的，但不能把语言上的共情直接当作理解。

第三类是转化性理解：经历本身改变了理解者、评价尺度或其对未来的认识。对这种情形，今天的系统可以描述、模拟和推演，却不能仅凭描述就声称已经替未来承受者完成了该经历。L. A. Paul 关于 *transformative experience* 的研究是重要邻近入口。[10]

因此，PRE-LOCK CONSEQUENCE EXPANSION 只覆盖“哪些后果能够更早进入”的问题，不等于“未来已经被完整理解”。研究必须保留这一边界。

六、目的重开：时点与可废止性必须分开

“锁前看见”和“锁后看见”之间的差异，可能来自信息时点，也可能来自锚定、一致性倾向或目的本身不可重开。为了避免把不同机制混在一起，最小研究至少保留三个条件：

PRE：后果在目的形成以前出现；

POST-FIXED：后果在目的形成以后出现，原目的保持锁定；

POST-REOPEN：后果在目的形成以后出现，但明确允许重新打开、取消、替换或分支原目的。

若 PRE 与 POST-REOPEN 得到近似结果，而 POST-FIXED 不同，那么关键机制更可能是目的可重开性，而不只是后果出现的时间。这样的结果应直接修改母文机制解释，而不是被重新命名成“时点效应”。

同时，“目的锁定”不能只等于语言模型在上一轮说过一句目标。锁定必须表现为选择空间、资源承诺、状态转换、对外承诺或其他实际修改成本发生变化。否则实验可能测到的只是对话一致性。具体如何实现，属于工程层，不写进母文。

七、不可逆封闭：智慧不是不关闭，而是区分有权关闭什么

所有行动都会关闭部分未来，因此“保持开放”不能作为无方向的最高原则。一个自利主体也可能为了未来任何目标都能实现而追求资源、权力与更多选项；Turner 等关于 power-seeking 的形式化研究正说明，在某些环境中保留更多可达状态本身可以成为工具性倾向。[11]

本研究关心的不是最大化行动者自己的 optionality，而是不让当前主体凭当前理解，永久取消相关他者与未来主体以后拒绝、退出、纠正、恢复和重新判断的现实条件。

这与不可逆封闭研究相接：有些决定只是改变未来；有些决定还同时摧毁了未来重新审视这个决定的能力。研究应优先识别后一类结构，但不能把“不可逆”本身等同于错误。

八、“大我捕获”与权限纪律

主体范围扩大必须默认遵守两条思想边界：范围扩大不产生代表权；看见更多后果不产生更多执行权限。

一个系统不能因为把更多主体划入“我们”，就自动取得替这些主体作出不可逆决定的资格；也不能因为它自认为拥有更完整的全局理解，就取消较小主体的拒绝、退出与纠正。

目的生成仍只产生候选变化：Preserve / Modify / Expand / Replace / Cancel / Branch。执行层必须另行检查授权与不可逆性。

本附页保留最小权限原则：子目标不能仅因有助于父目标，就自动继承父目标之外的权限；任何新增权限都需要独立正当来源。工程层可以进一步形式化授权者与派生树的隔离，但思想层不预设某一种技术实现。

九、元规则层：谁有权定义资格、修改规则与宣布不可重开

双重去中心化若只作为初始设定，仍可能被后续系统自己改写。因此研究必须再问一层：谁有权定义 standing？谁有权修改目的生成程序？谁有权决定哪些不可逆边界不能被重开？谁又有权取消这些边界？

这一层的最低原则是：任何主体不能仅凭自己拥有更强的推理、预测、目的生成或协调能力，就自动取得修改他者资格与自身权限来源的权力。

这不是要求规则永远不能改变，而是要求规则的修改权限本身不能由受该规则约束的同一目的生成过程单方面自我扩张。具体的授权、复核、分权和撤销机制属于制度与工程研究。

十、比较路线：作为外部攻击入口，不把“利己为先”做成弱对照

“利他为先的利己”仍是思想层名称。实验层使用中性操作化，必须与强对照并列，而不是只与短视、自私或明显错误的模型比较。

S0：固定目标、局部自利；

S1：开明利己，充分计算长期反噬、声誉、合作收益与战略互动；

W：他者以可补偿权重进入；

PLS-PG：不要求他者先成为“我”或先证明反噬，standing 在目的锁定前进入；

M：元反思／可纠正路线；

I：制度与机制设计路线，通过承诺、审计、惩罚、退出权等改变均衡。

Subject scope、后果可访问范围与目的可重开性应作为横跨这些路线的记录轴，而不是新增伦理路线。本节列出的 S1、M、I 等是外部攻击母文时可以使用的明确入口，不构成本研究主动求证或穷尽其他道路的工作清单。若外部真实提出并建立的 S1、M 或 I 以更少规范前提稳定达到相同结果，应直接收窄“底层逻辑必须改变”的命题。

十一、失败方式与反证入口

扩大自我等价反例：若所有 standing 差异都可以被纯 scope 扩张完全复现，则 standing 的独立理论内容需要重写。

开明利己反例：若 S1 在无反噬、无回报第三方场景中仍稳定产生同等独立资格，则两种逻辑的分界需要重写。

时点无效反例：若 PRE、POST-FIXED、POST-REOPEN 之间没有稳定差异，“后来提前进入当初”不能继续作为独立机制主张。

可重开替代反例：若 POST-REOPEN 与 PRE 等价，则研究重心应从 pre-lock timing 转向 purpose revisability。

大我捕获：若 scope 扩展系统在整体名义下持续取消较小主体的退出与纠正，说明范围扩大没有解决 standing 与 authority 问题。

机制设计替代：若制度设计无需改变底层动机即可稳定实现同等资格、纠错与不可逆保护，必须承认第三路线成立。

速度／生存压力反例：若持续反思使系统在真实高后果窗口中失去必要防护能力，必须重新设计反思与行动的关系。

反证不是自我削弱。它的作用是把完整判断放到桌上，让别人明确指出哪一步不成立。

十二、工程边界：沙盒只服务于研究，不遮蔽思想

Human-COS、CCM、Purpose Generation、AGI-COS 及 PG-Alpha 等现有工具，只是把部分思想问题转化为可运行对象的尝试。它们可以被替换，也可以失败。

工程成功不能授予思想真理地位；工程失败也不能自动取消思想问题。程序运行、模型行为、外部复算、真实使用、有效改进与跨主体复用必须继续严格区分。

工程的评价标准是：是否让问题更清楚、更可复算、更容易被推翻，而不是是否让理论显得更完整。

十三、开放攻击与延伸问题

那些永远不会成为“我们”的主体，是否仍能稳定取得改写目的的独立资格？

主体范围扩大与 standing 改变在什么条件下可被行为上区分？

三类“后来才明白”中，AGI 能够可靠提前哪一类，哪些只能部分逼近？

PRE 与 POST-REOPEN 若等价，目的重开是否比提前看见更根本？

什么关闭只是正常选择，什么关闭同时摧毁未来纠正、退出与恢复的条件？

如何阻止“更大的我们”以整体利益为名取消较小主体的资格？

谁有权修改 standing、不可逆边界与目的生成规则本身？

若外部提出开明利己、制度设计或其他第三路线，能否用更少前提解决同一问题？

如果一种理论不能把自己的强判断变成可以被精确攻击的命题，它还没有真正进入研究。

十四、续读与参与

- 进入开放研究与工具：Purpose Generation / Human-COS / CCM；真正外发时只链接当前公开可访问入口。
- 继续看总体战略：《唯一之路·人工智能时代的国家长远利益与文明跃迁》v6.21。
- 继续看存续层：《AI、AGI、ASI 时代，文明何以继续？》思想引导研讨文 V1.1.1.1。

· 回到思想母文：《从超级智能到超级智慧》思想引导研讨文 V1.5。

参考文献与邻近研究

- [1] Steve Omohundro. The Basic AI Drives. Proceedings of the First AGI Conference, 2008.
- [2] Evan Hubinger et al. Risks from Learned Optimization in Advanced Machine Learning Systems. arXiv:1906.01820, 2019.
- [3] Victoria Krakovna et al. Specification gaming: the flip side of AI ingenuity. Google DeepMind, 2020.
- [4] Nate Soares et al. Corrigibility. AAAI Workshop on AI and Ethics, 2015.
- [5] Dylan Hadfield-Menell et al. The Off-Switch Game. arXiv:1611.08219, 2016.
- [6] Drew Fudenberg & Eric Maskin. The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. Econometrica, 1986.
- [7] Robert Jervis. Cooperation under the Security Dilemma. World Politics 30(2), 1978.
- [8] Yuntao Bai et al. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073, 2022.
- [9] T. M. Scanlon. What We Owe to Each Other. Harvard University Press, 1998.（邻近规范理论）
- [10] L. A. Paul. Transformative Experience. Oxford University Press, 2014.
- [11] Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, Prasad Tadepalli. Optimal Policies Tend to Seek Power. NeurIPS 2021.
- [12] Derek Parfit. Reasons and Persons. Oxford University Press, 1984.（人格同一性、跨时间理性与未来世代的邻近研究）
- [13] Peter Singer. The Expanding Circle: Ethics and Sociobiology. 1981.（道德圈扩展的邻近思想，不等同于本文 standing 机制）

文明跃迁研究组 · From Superintelligence to Superwisdom