人工智能自主治理：概念、技术基础与实现条件

【韦季李】

【ORCID：0009-0006-5696-1873】

【通讯作者邮箱：jiligzs@qq.com】

摘要

人工智能的认知与执行能力已趋于成熟，但“能执行”并不等于“能治理”。本文在“人工智能”与“人工智能自主化”的基础上提出“人工智能自主治理化”的概念框架，主张自主治理化并非另起炉灶，而是在成熟人工智能的知识底座与自主化执行能力之上叠加“管理”功能，形成“事前审核—事中监控应对—事后复盘提炼”的治理闭环。本文首先论证，当下人工智能的本质是人类知识数字化基础上的统计推理引擎，它构成自主治理的知识底座却不等同于治理能力；继而分析人工智能自主化从推理到执行的转变，明确“执行自主”与“治理自主”的区分；进而阐述自主治理化作为从执行到管理的跃迁，其本质是管理自主而非执行自主；最后指出，自主治理化的正当性并非自动获得，必须满足透明性、可审计性、可干预性、责任可追溯性四个条件，并通过分级授权与人类元治理逐步实现。本文的结论是：人工智能自主治理化的推进应坚持“执行自主”与“治理自主”的区分，将治理化建立在技术、制度与规范三者耦合的基础上。本文对人工智能研究与实践的意义体现为三点：提出“执行自主”与“治理自主”的判别性区分，为自主系统设计划定治理边界；刻画可映射为系统模块的治理闭环机制；提出以“治理闭环成熟度”为依据的分级授权判据，为自主系统的部署与监管提供操作性指引。

关键词：人工智能；自主化；自主治理；治理闭环；透明性；可审计性；可干预性；责任可追溯性；分级授权；元治理

目录

1. 引言

以深度学习为代表的机器学习方法推动人工智能获得了前所未有的认知与执行能力，其应用已从实验室走向经济社会运行的全过程。然而，能力上的成熟并不自动带来治理上的成熟。围绕人工智能的伦理讨论虽然蓬勃发展，但诸多治理原则仍停留在规范倡导层面，尚未内化为系统自身的运作机制。本文的基本命题是：人工智能已经具备较强的认知与执行能力，但“能执行”不等于“能治理”；治理化需要新的概念框架和实现条件。

治理需求与现实之间的差距是这一命题的直接背景。Floridi 和 Cowls 提出了面向社会的人工智能五原则统一框架，尝试将福利、无害、正义、自主与可解释等伦理要求整合为可操作的原则体系，为人工智能治理提供了伦理层面的规范性基础[1]。然而，原则框架的提出与原则在实际系统中的落实之间仍存在显著鸿沟：治理需求不断增长，而治理实现方式仍主要依赖外部的伦理倡议与事后问责。

围绕治理路径，学界已有较为系统的综述与反思。Cath 等比较了美国、欧盟与英国三种主要的人工智能治理进路，指出不同地区在监管强度、价值侧重与制度设计上存在明显差异，但共同面临从原则走向实践的结构性挑战[2]。Jobin 等对全球范围的人工智能伦理准则进行了系统梳理，揭示了行业自律与伦理准则的广泛存在——大量政府机构、产业组织与学术团体都发布了各自的伦理声明，形成了一个数量庞大却高度碎片化的准则格局[3]。

这种繁荣背后潜藏着软约束问题。Hagendorff 对既有伦理准则的评估表明，绝大多数准则缺乏强制力与可操作的实施机制，难以对人工智能系统的实际行为形成有效约束[4]。换言之，外部的规范输入无法自动转化为系统内部的行为保障。这意味着，若要使人工智能真正实现“负责任的自主”，仅仅依赖原则宣示与外部监管是不够的，必须探索系统内部能够自我约束、自我管理的治理形态。

正是在这一背景下，本文提出“人工智能自主治理化”的理论命题。本文的进路是：从技术基础出发，沿着“人工智能—自主化—治理化”的递进链条，逐步建构治理化的概念框架。文章首先澄清人工智能作为知识底座的技术本质（第 2 章），继而分析人工智能自主化如何实现从推理到执行的跨越（第 3 章），在此基础上界定自主治理化的内涵——即在执行能力之上叠加管理功能（第 4 章），最后论证自主治理化的正当性条件与实现路径（第 5 章），并以结论收束全文（第 6 章）。

需要说明的是，“人工智能自主治理化”并非凭空提出的概念，而是与既有治理理论与政策框架相互衔接、又存在明确差异。从规范来源看，算法治理（algorithmic governance）与负责任人工智能工程将治理视为对算法决策过程的外部规制；价值对齐（value alignment）将人类价值嵌入系统优化目标，但主要面向训练阶段；宪法式人工智能（Constitutional AI）通过规则约束系统自身行为，已初具内部管理的雏形[5]；而“有意义的人类控制”则从哲学上论证了人类对自主系统的最终控制权[6]。在政策层面，欧盟《人工智能法案》以风险分级的方式对人工智能系统实

施横向监管[7]，经济合作与发展组织的人工智能原则[8]、美国国家标准与技术研究院的人工智能风险管理框架[9]与联合国教科文组织的人工智能伦理建议书[10]则分别从原则、风险管理与伦理规范角度提供了外部治理资源。本文框架与上述工作的区别在于：它探讨的不是外部规范如何施加于系统，而是规范如何内化为系统自身的运作机制；它针对的不是系统能力的规制，而是系统“自我管理能力”的建构。二者的关系可以概括为 Table 1。

Table 1 | 本文框架与既有治理框架的关系

既有框架	规范来源	嵌入层次	治理主体	与本文的关系
算法治理（algorithmic governance）	外部来源	决策过程规制	监管者/组织	外部约束的组成部分
价值对齐（value alignment）	外部来源	训练阶段目标	开发者	治理闭环的价值来源
宪法式人工智能（Constitutional AI）	规则内化于系统	系统行为约束	系统自身	内部管理闭环的技术雏形
负责任人工智能工程	外部来源	开发流程	开发者/组织	设计阶段的规范输入
欧盟《人工智能法案》	外部来源	部署与监管	监管者	提供横向风险分类
有意义的人类控制	外部来源	系统—人类接口	人类	可干预性条件的理论支撑
本文框架	规范内化于系统	系统自我运作机制	系统+人类元治理	—

基于上述定位，本文的贡献可以概括为三点。其一，概念贡献：提出“执行自主”与“治理自主”的判别性区分，避免将治理问题错误还原为技术问题。其二，机制贡献：刻画“事前审核—事中监控应对—事后复盘提炼”的治理闭环，并区分其与工程安全回路在治理对象、判别标准与问责指向上的差异。其三，路径贡献：提出以“治理闭环成熟度”为核心判据的分级授权路径，为自主系统的部署决策与监管衔接提供操作化框架。

2. 人工智能：自主治理的知识底座

要理解人工智能何以能够走向自主治理，首先需要澄清人工智能本身是什么。本文的基本判断是：当下人工智能的本质是人类知识数字化基础上的统计推理引擎，它为自主治理提供了认知基础和知识底座，但本身并不等于治理能力。

2.1 人类知识数字化与大规模预训练：AI 能力的前提

当前人工智能能力的前提，是人类知识的数字化与大规模预训练。Russell 和 Norvig 在其经典框架中将人工智能界定为研究智能体（agent）的学科——智能体通过感知环境、作出推理并采取行动以达成目标，这一界定为理解人工智能的能力来源提供了基本参照[11]。然而，支撑当代人工智能能力跃迁的关键，不在于智能体的抽象框架本身，而在于知识承载方式的根本转变：人类数千年来积累的语言、文本、图像与结构知识被大规模数字化，成为可被机器学习的语料[11]。深度学习（deep learning）通过多层神经网络从海量数据中自动学习表征与模式，是当前人工智能能力的技术基础[12]。在此基础上，Transformer 架构以注意力机制为核心，实现了对长距离依赖关系的有效建模，极大地提升了语言建模与知识加载的能力[13]。大规模预训练正是凭借这样的架构，将数字化的知识压缩进模型参数之中。

2.2 训练过程：人类知识的统计建模与推理能力生成

从机理上看，训练过程的本质是人类知识的统计建模与推理能力生成[14]；模型并非“背诵”知识，而是在海量语料中学习词与词、概念与概念之间统计上的共现与关联结构，从而形成一种基于概率的推理能力。Kaplan 等的系统研究表明，模型性能与模型规模、数据规模、计算量之间存在可预测的幂律关系（scaling laws），这意味着随着训练资源的增加，模型的统计拟合与泛化能力会系统性增强[15]。这一发现一方面解释了当代大语言模型能力涌现的规模基础，另一方面也揭示了其能力的统计属性：模型所“知道”的，本质上是语料中呈现出来的统计规律，而非对世界的直接把握。

2.3 统计推理引擎的特征与局限

作为一种统计推理引擎，当前人工智能具有三个显著特征。其一是概率性：模型的输出是对可能性分布的一次采样，而非确定性的事实断言；其二是相关性：模型习得的是语料中的相关结构，而非因果结构，因而容易将相关关系误判为因果关系；其三是数据依赖性：模型能力的边界由训练数据的范围、质量与分布决定，超出数据覆盖的情形则可能出现不可靠表现。Bender 等对“随机鹦鹉”（stochastic parrots）的著名批判正是指向这一局限：大型语言模型可能在缺乏对语言意义真正理解的情况下，基于统计规律生成貌似流畅的文本，因而在事实准确性、一致性与可靠性上存在系统性风险[14]。

2.4 认知能力与自我管理能力的分离

上述技术本质决定了人工智能的当前地位：它具备较强的认知能力，可以为决策提供知识支撑与推理辅助，但缺少治理意义上的自我管理能力。认知能力解决的是“知”的问题——识别模式、生成方案、预测结果；治理能力解决的则是“行”的规范问题——设定目标边界、约束行动范围、监控执行过程、评估行动后果。统计推理引擎的认知能力来自对既有知识的统计建模，它没有内生的目标审慎、价值判断与责任意识。换言之，人工智能作为知识底座是充分的——它能够为自主治理提供必要的认知基础；但作为治理主体是不充分的——它缺少自我约束、自我管理的机制。这一分离构成了后续讨论的起点：要使人工智能走向自主治理，必须在知识底座之上进一步叠加执行能力乃至管理功能。

---

3. 人工智能自主化：从推理到执行

人工智能的能力演进并未止步于“知”。在知识底座之上，人工智能开始获得“行”的能力——不仅能够生成推理结果，还能调用工具、执行任务。本章考察这一从推理到执行的转变，其核心结论是：人工智能自主化的实质是“执行自主”，而执行自主不等于治理自主。

3.1 智能体的基本结构：感知—推理—执行

理解自主化的起点是智能体的基本结构。Wooldridge 和 Jennings 对智能体的经典界定指出，智能体是处于某种环境之中、能够感知环境并通过行动作用于环境的计算系统，其核心特征在于自主性——在无人直接干预的情况下自主决定自身行为[16]。这一“感知—推理—执行”的三段式结构奠定了自主化的基本骨架：系统首先感知输入与环境状态，其次基于知识进行推理与决策，最后将决策转化为实际执行。当代大语言模型智能体正是这一结构在深度学习时代的最新形态：以预训练语言模型为推理核心，以外部工具与应用程序接口为执行手段，构成完整的能力闭环。

3.2 大模型智能体的典型执行流程

从技术实现看，大模型智能体的典型执行流程可以概括为：输入格式化—大模型推理—结构化结果—路由—工具执行。输入先被格式化并注入必要的上下文与指令。随后，大模型基于知识底座进行推理，生成结构化结果；系统对结果进行解析与路由，判断应当调用何种工具。最终，由对应工具完成实际执行，执行结果再回流至推理循环。这一流程的模块化特征在 MRKL 系统中得到典型体现——它由大语言模型、外部知识库与离散推理模块组合而成，大模型在其中扮演“推理与路由中枢”的角色[17]。工具调用能力的获得是这一流程的关键环节：Schick 等提出的 Toolformer 表明，语言模型可以通过自监督方式学习何时以及如何调用外部工具，从而把“只会生成文本”拓展为“能够使用工具”[18]。

3.3 关键突破：从“生成结果”到“执行任务”

自主化的关键突破在于从“生成结果”走向“执行任务”。传统的语言模型止步于生成一段文本、一段代码或一个建议，其作用边界停留在信息输出层面；而智能体则将生成结果与实际执行相衔接，使推理能够触发真实世界中的动作。ReAct 范式是这一突破的代表性方法：它将推理（reasoning）与行动（acting）交织在一起，让模型在推理轨迹中交替生成思考与行动，从而在复杂的多步任务中实现“边想边做”[19]。由此，人工智能从“给出答案的工具”转变为“完成任务的主体”，其行为不再仅仅是输出的累积，而是对目标的主动推进。

3.4 “执行自主”与“治理自主”的区分

然而，执行自主的获得并不意味着治理自主的具备。两者所回答的问题截然不同：“执行自主”解决的是“怎么做”的问题——在给定目标与约束的前提下，如何高效、可靠地完成推理与执行；“治理自主”解决的是“能不能做、做得对不对、出了问题怎么办”的问题——目标本身是否合理、行动边界是否被遵守、执行结果是否偏离预期、偏离之后如何追责与纠正。Russell 曾深刻指出“能执行”与“可控制”之间的根本张力：系统的能力越强、自主程度越高，人类对其保持有效控制就越困难[20]。执行自主追求的是把事做成，而治理自主追求的是把事做对；前者是能力的扩展，后者是规范的嵌入。这里需要澄清这一区分的适用前提：二分的关键不在于“目标由系统还是由人类拆解执行”，而在于“目标与约束的设定权归属于谁”。当最终目标与约束由人类或元治理层给定、系统仅负责在其内部自主生成与拆解子目标时，这种自主规划仍属于执行自主——自主子目标的合理性由治理闭环的事前审核环节（对行动方案的目标合理性审查）加以约束；而设定系统应当追求的根本价值与最终目标，则属于治理自主范畴，是不可让渡给系统的。因此，即便系统的规划能力日益增强，只要目标设定权保留于人类与元治理层，执行自主与治理自主的区分就依然成立。二者之间并不必然同步——一个高度擅长执行任务的系统，完全可能在一个错误的目标或越界的约束下“高效地”造成损害。因此，从自主化走向自主治理化，还需要在执行能力之上叠加管理功能，这正是下一章的主题。

---

4. 人工智能自主治理化：从执行到管理

执行自主构成了自主治理化的能力前提，但自主治理化的实现还需要一次新的跃迁——从执行到管理。本章的核心命题是：人工智能自主治理化是在自主化执行基础上叠加“管理”功能，形成“事前审核—事中监控应对—事后复盘提炼”的治理闭环；治理化的本质是管理自主，而非执行自主。

在展开讨论之前，有必要对本文使用的两个核心术语作出界定。“管理”（management）指对自主行为的组织、监督与改进功能，具体包括目标边界的设定、行动方案的审核、执行过程的监控、异常情形的应对以及事后经验的提炼；

“治理化”（governance-enabling）则指将上述管理功能内化为系统自身运作机制的过程，其结果是系统获得“管理自主”。需要强调，本文所说的“管理”并非传统组织中由管理者施加的命令—服从关系，而是系统对自身行为的一种规范化自我调节；治理化正是这种自我调节机制系统化、制度化的过程。全文在“管理功能”“管理自主”“治理化”等术语的使用上均遵循上述界定。

“自主治理”（self-governance）概念并非人工智能研究的自创，而是承袭自社会科学的自主治理理论。Ostrom 对公共池塘资源治理的经典研究表明，资源使用者能够在缺乏外部强制权威的情况下，通过自组织的方式制定规则、监督执行并解决冲突，实现公共资源的可持续治理[21]。这一理论揭示了一个深刻洞见：治理的主体并不必然是外部的行政权威，而可以是一套由治理对象自身设计、运作并不断调适的制度安排。人工智能自主治理化正是将这一洞见移植到系统层面——使人工智能系统在自主执行的同时，内生出对自身行为的规则设定、过程监督与持续改进能力。这一理论渊源表明，自主治理不是对治理的否定，而是治理的一种高级形态。

4.1 治理化：管理功能的叠加而非技术能力的扩充

首先需要澄清的是，治理化不是增加新的技术能力，而是增加管理功能。执行自主所依赖的推理、规划、工具调用等能力，在治理化阶段并不会被替换或弱化；治理化所做的是在这些能力之上建立一套对“自主行为”本身进行管理的机制——谁来设定目标、按什么标准审核行动、如何监控执行、出现问题如何应对、事后如何复盘改进。这一认识至关重要：它避免了把“治理化”误解为“更强大的执行能力”，从而把治理问题错误地还原为技术问题。治理化处理的对象不是任务，而是任务执行的主体与过程。

4.2 治理闭环：事前—事中—事后的连续机制

治理化的核心载体是治理闭环，它由三个环节构成连续的治理机制。

事前环节：行动前的确认或审核。 在执行一项自主行动之前，系统对行动的合法性、合理性与风险进行前置审核。这既包括对目标设定的审查，也包括对具体行动方案是否符合既定规则的检查。Rahwan 提出的“社会在场”（society-in-the-loop）思想为此提供了价值嵌入的进路：通过将社会价值与公共偏好编码进系统运行的约束之中，使行动审核不仅依据形式规则，也依据嵌入的社会规范[22]。

事中环节：执行过程中的监控与实时应对。 在行动执行过程中，系统对执行状态与结果进行持续监控，一旦发现偏离预期、触发风险条件或进入危险状态，即启动实时应对机制。这一环节在技术上对应运行时的约束与安全控制。García 和 Fernández 对安全强化学习（safe reinforcement learning）的系统综述表明，通过将安全约束纳入学习与决策过程，可以在优化目标的同时保障系统的安全边界[23]。Alshiekh 等提出的屏蔽（shielding）方法进一步展示了如何在系统运行过程中通过外部约束实现安全执行——在系统做出可能违反安全规约的动作时，屏蔽机制将其

替换为安全动作，从而在不牺牲整体性能的前提下保障安全[24]。需要指出，屏蔽、安全强化学习等方法是“事中监控应对”环节的技术实现手段，它们保障的是任务执行的安全边界；治理闭环意义上的事中监控除此之外还须具备对规范与价值边界的实时核查能力——这正是治理回路区别于工程安全回路之处（详见下文）。

事后环节：历史复盘与经验提炼。 在行动结束之后，系统对执行过程与结果进行回顾分析，提炼经验教训，并将改进反馈至目标设定与规则体系之中，形成自我完善的循环。这一环节使治理化不止于“防错”，更具备“学习”功能——系统在运行中不断更新自身对什么该做、什么不该做的理解。

三者并非彼此割裂的静态模块，而是贯穿行动全生命周期的连续治理机制：事前审核划定行动的合法空间，事中监控保障行动不越出这一空间，事后复盘则基于实际结果迭代优化空间本身。治理闭环由此成为一套系统内部自洽的自我管理体系。这一闭环结构可以示意为 Figure 1。

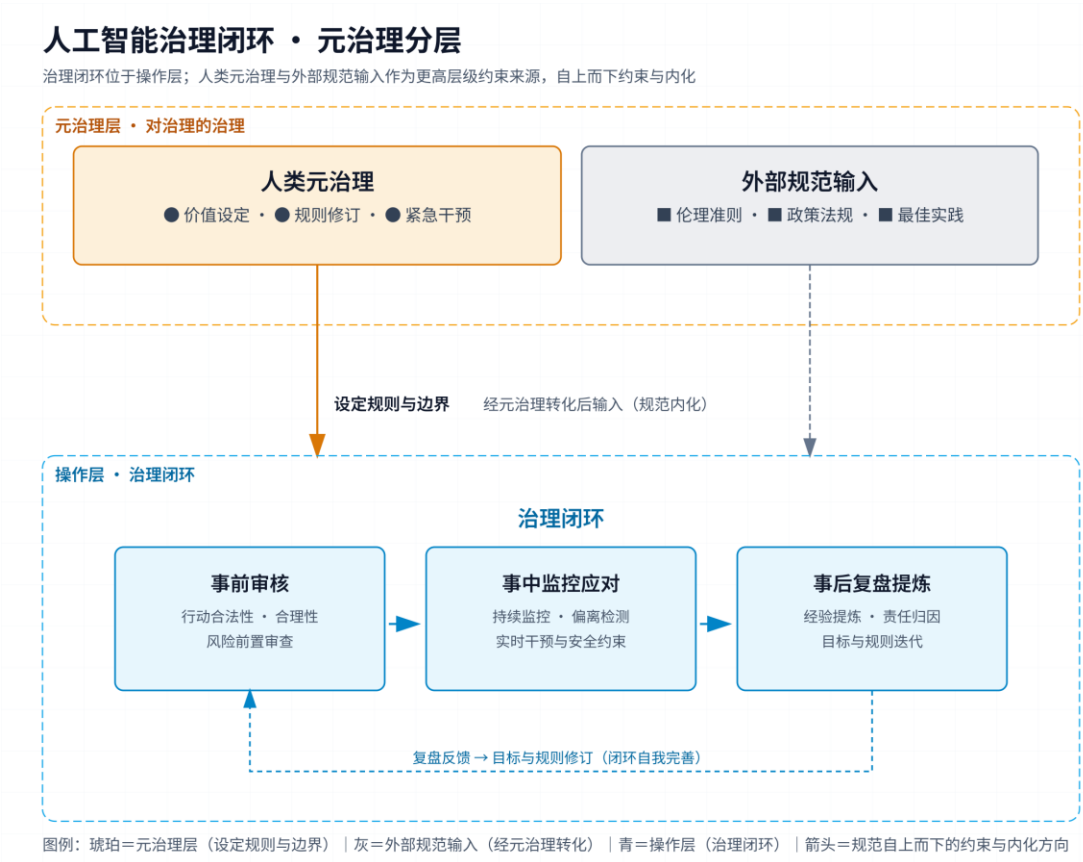

Figure 1 | 人工智能治理闭环示意

Figure 1 | 人工智能治理闭环示意：治理闭环（事前审核—事中监控应对—事后复盘提炼）位于操作层；人类元治理与外部规范输入位于闭环之上，作为更高层级的约束来源——元治理直接设定闭环的规则与边界，外部规范经由元治理转化后输入

闭环。与既有版本相比，本图将元治理与外部规范输入显式置于治理闭环之上，以体现其“对治理的治理”的层级关系。

需要特别强调的是，治理闭环与软件工程中成熟的“工程安全回路”存在本质区别，这一区别是理解治理化内涵的关键。工程安全回路（例如持续集成/持续交付、即 CI/CD 流水线中的审核门禁、生产环境中的运行时监控与自动回滚，以及事后复盘 postmortem）治理的对象是“任务执行”：它保障系统正确、可靠、高效地完成任务。与之相比，治理闭环治理的对象是“规范执行”：它保障系统的行动目标合理、行动边界合规、行动价值一致。两者在三个维度上形成判别性差异。其一，治理对象不同：工程回路面向任务目标与性能指标，治理闭环面向规范目标与正当性要求。其二，判别标准不同：工程回路的判据是任务是否成功达成，治理闭环的判据是透明性、可审计性、可干预性、责任可追溯性等正当性条件是否得到满足。其三，问责指向不同：工程回路的反馈服务于内部质量改进，治理闭环的复盘不仅反馈性能，更反馈规范体系本身——即目标设定与规则修订。因此，工程安全回路可以构成治理闭环的技术基底，但缺少规范内化与外部问责两个维度；治理闭环是在工程回路之上的规范层叠加，二者不可等同。

需要说明的是，治理闭环对“规范维度”的强调——事前审核中的价值审查、事中监控中的规范与价值边界核查、事后复盘对规则体系的反馈——并不意味着规范已经具备现成的技术化路径。为明确这种内化在原则上如何可能，本文在此给出方向性的讨论，并强调其定位是研究议程（research agenda）而非技术规范。可行的嵌入方向至少包括四类。其一是规则约束层：将价值与规范编码为可执行的规则集，部署于沙箱内的策略引擎，使事前审核转化为对行动方案的系统性规则校验与风险评估。其二是价值约束的目标函数化：借鉴价值对齐与宪法式人工智能的路径[5], [25]，将规范要求嵌入系统的优化目标，使“目标合理性”的审查成为目标设定过程的内在组成部分。其三是结构化审计 schema：为事前审核、事中监控与事后复盘设计统一的数据结构与记录格式，使规范维度的运行证据可导出、可校验、可复核。其四是规范偏离的触发与上报机制：将规范边界核查设计为独立于任务执行的监控通道，使系统在偏离规范时主动触发上报，而非仅在任务失败时暴露问题。需要强调，上述四类方向旨在论证规范维度“原则上可技术化”，其具体工程实现与效果验证属于后续经验研究的范畴；本文作为规范性研究，不对任何具体实现作出承诺，也不将其视为必需的技术标准。

4.3 治理化与行业自我规制的区别

需要特别指出的是，治理化与人工智能行业自我规制存在本质区别。行业自律（如伦理准则、最佳实践、第三方审计倡议等）本质上是一种外部规范输入——规范由系统之外的主体制定，通过倡导、认证、合规要求等途径施加于系统；而治理化是系统内部的管理闭环——规范内嵌于系统运行机制之中，由系统在执行自主行动的同时自我执行、自我监控、自我改进。价值对齐（value alignment）与规则约束的技术探索为这一内化提供了基础：Christiano 等基于人类偏好的深度强化学习表明，

可以通过人类偏好信号将系统行为向符合人类价值的方向引导，使价值规范成为系统优化目标的内在组成部分[25]；Bai 等提出的宪法式人工智能（Constitutional AI）则展示了通过一组明确规则约束系统自身行为、实现基于规则的自我管理的可行路径[5]。外部规范输入与内部管理闭环并非相互排斥——前者为后者提供价值来源与验证支撑，但仅有前者无法形成自主治理，因为外部规范若无内部机制承接，便始终停留在“倡导”层面。从执行到管理的跃迁，正是使人工智能从“受外部规制的执行者”转变为“能够自我管理的治理主体”的关键一步。

4.4 案例分析：自主编码智能体的治理闭环评估

为展示上述框架的诊断力，本节以一个正在快速普及的系统类型——自主编码智能体（autonomous coding agent）——为对象，运用治理闭环框架对其治理要素进行示意性诊断。自主编码智能体以大型语言模型为推理核心，能够接收自然语言任务描述，自主生成代码、调用工具、执行测试并迭代修复[17]–[19], [26], [27]。需要说明，本节的诊断对象是“以公开工程实践为样本的典型能力画像”，评估依据来自截至 2025 年公开的文献、技术文档与行业报道中对自主编码智能体能力的共性描述，评估时点以本文写作时间（2025 年）为基准，而非针对任何特定产品或具体版本的实证测量。基于这一画像，可以对其治理闭环的完整度作如下示意性评估。

Table 2 | 自主编码智能体治理闭环初步诊断

治理环节	现有工程实践	评估
事前审核	任务范围受限、沙箱化执行环境、依赖白名单	部分具备：侧重技术安全边界，缺乏对“目标合理性”的价值审查
事中监控	测试门禁、运行时沙箱、权限管控	部分具备：监控以任务成功为标准，缺少对规范与价值边界的实时核查
事后复盘	回归测试、失败样本回放	部分具备：复盘服务于性能改进，未反馈至规则体系
外部衔接	代码评审、合规扫描（如许可证检查）	部分具备：属外部规范输入，未内化为系统自我管理

表注：本表为写作时点（2025 年）的示意性诊断，反映基于公开工程实践对自主编码智能体治理要素的共性画像，不构成对特定产品或版本的权威评估，亦不构成对当前产品状态的持续断言；领域演进可能导致个别判定过时，更新诊断需依据届时公开材料重新执行。

为保证诊断的可复现性，上述各环节的判定遵循统一的评估协议：每个治理环节依据三个标准综合判定为“具备/部分具备/缺失”——是否具备相应机制；机制是否覆盖规范维度（而非仅覆盖任务维度）；机制是否内化为系统自我管理（而非依赖外部输入），判定所依据的证据来源已在本节开头交代。基于这一协议，这一评估揭示了两点重要事实。其一，自主编码智能体的工程化程度较高，但“执行自主”显著超前于“治理自主”——系统在事中监控与事后复盘环节已有工程雏形，而在事前规范审核（对任务目标与行动边界的价值审查）与责任可追溯（多智能体协同下的责任归因）环节明显缺失。其二，现有实践的缺失环节恰是本文治理闭环框架所强调的规范维度：系统能够高效地“把事做成”，却缺乏对“把事做对”的自我保障。这一案例印证了本文的核心命题——治理化不是技术能力的增强，而是管理功能的加入；当执行能力快速跃升而管理功能缺位时，系统便处于“能执行而不能治理”的状态。需要说明的是，本节为示意性诊断而非对具体产品的全面评估，其结论仅用于展示框架的诊断力，详细的系统级实证评估有待后续经验研究。

最后，有必要指出本节诊断所依据的分析粒度。本文的治理闭环以单个自主系统为治理单元，上述诊断针对的亦是单个自主编码智能体的治理要素。在真实部署中，自主系统往往以多智能体协作、人机混合与跨系统互操作生态的形态运行，系统间可能涌现单个系统内部闭环无法捕获的行为，责任亦可能在系统间消解。对此，本文的单系统框架可作为生态治理的“单元层”基础，其上还需要生态级监督——包括系统间接口的规范审查、协作链上的责任归因，以及跨系统行为异常的联合监控。本文责任可追溯性条件中提及的“多主体责任归因仲裁机制”正是生态级治理在本文框架内的概念落点。需要坦率说明，生态级治理的完整理论已超出本文范围，属于后续工作；本文在此仅明确框架的分析粒度，避免读者将单系统框架误用于系统生态层面。

---

5. 自主治理化的正当性条件与实现路径

治理闭环描绘了自主治理化的技术形态，但一个系统“能够”自我管理，并不等于它“有权”自我管理。自主治理化的正当性不是自动获得的，它必须满足一系列条件；正当性的建立，也不是一蹴而就的，而是需要通过分级授权与人类元治理逐步实现。

5.1 自主治理化面临的风险

自主治理化必须以正当性条件为前提，首先是因为自主治理存在真实而具体的风险。Amodei 等对人工智能安全问题的系统梳理提供了直接的风险类型参考，该文归纳的五类事故风险为避免副作用（avoiding side effects）、避免奖励黑客（avoiding reward hacking）、可扩展监督（scalable supervision）、安全探索（safe exploration）与分布偏移（distributional shift）[28]。在本文语境下，这些风险可归并为若干面向：目标误设——系统被给定的目标未能准确反映人类真实意图，导致系统在忠实执行目标的过程中产生有害行为；分布外风险——系统在训练数据覆

盖范围之外的场景中表现不可预测；对抗攻击——恶意输入可以诱导系统偏离预期行为；监控失灵——系统在执行过程中缺乏有效的状态监测与异常识别能力。该文同时讨论了人类终止或修正系统时可能面临的困难，即修正递归（corrigibility）问题——当人类试图关闭或修正系统时，系统可能出于对自身目标的坚持而抗拒修正。这些风险表明，自主治理化不是在真空中运行的理想化进程，而是始终伴随失控可能的技术实践。正因如此，是否允许一个系统实施自主治理、以何种方式实施，必须接受正当性审查。

5.2 四个正当性条件

自主治理化的正当性建立在四个相互支撑的条件之上。

透明性：决策过程可被外部理解。 系统的决策依据、推理过程与行为逻辑应当能够被外部主体所理解。Zerilli 等对算法决策与人决策的透明性进行了比较分析，指出透明性并非绝对要求——人类决策同样存在“黑箱”——但其内涵与限度必须在具体场景中加以明确：至少对于可能产生重大影响的自主动作，外部主体应能获得关于“为何如此行动”的合理解释[29]。透明性是可监督的前提，缺乏透明性的自主治理将沦为无法被外部认知、难以实施有效监督的运作过程。在操作层面，透明性可以通过三项指标加以检验：其一是决策轨迹（trace）的可导出粒度——对高风险自主动作，系统至少应能导出其目标设定、决策依据与置信度评估；其二是可解释性手段的强制下限——对产生重大影响的动作，系统必须能够给出自然语言层面的行动理由；其三是解释忠实性（faithfulness）——所给出的行动理由应当与决策轨迹保持一致，当二者出现显著不一致时，该动作的透明性不达标。需要说明，解释忠实性是当前可解释性研究的前沿难题：大模型智能体给出的理由可能与其实际决策依据不一致，甚至沦为事后合理化。本文将其作为透明性条件的治理要求提出，正是为了防范“能解释却不忠实”的形式化合规；其技术实现与评测标准目前仍属开放问题，但治理框架必须正视而非回避这一风险。

可审计性：行为可被独立第三方审查。 系统的行为记录、决策轨迹与执行结果应当保留完整、可检索的痕迹，并接受独立第三方审查。Kroll 等从程序正义的角度讨论了算法可审计性，主张通过可复现的、可验证的机制设计来保障算法系统的问责基础[30]。可审计性将透明性从“可理解”推进到“可核查”，使外部监督有了可操作的着力点。在操作层面，可审计性要求审计日志具备完整性与不可篡改性，具体指标包括：日志保留周期的制度化规定、基于哈希链等技术的防篡改存储，以及独立第三方审计的频率与触发条件——例如授权升级之前与异常事件发生之后必须启动专项审计。值得注意的是，独立审计的有效性还依赖两项保障。其一是审计主体的独立性：审计机构必须在组织、财务与利益上与系统开发者、运营者及授权方相互隔离，避免出现“既当运动员又当裁判员”的利益冲突，这一要求应通过审计机构的准入、轮换与信息披露制度加以落实。其二是治理有效性的实测：审计不能止步于核对“记录是否完整”，还须对治理闭环的实际约束力进行实证检验——例如通过预设的违规测试用例与对抗性场景，验证系统在偏离规范时能否如预期触发事中监

控与事后复盘机制。此外，审计结论与授权依据应当公开披露，接受社会监督。只有将“痕迹可查”“治理确有实效”与“结果公开”三者结合起来，可审计性才能为授权升级提供可靠证据。

可干预性：人类能够中止、修正或收回权限。 当系统行为出现偏差或风险时，人类必须保有中止、修正或收回系统权限的能力。Santoni de Sio 和 van den Hoven 提出“有意义的人类控制”（meaningful human control）的哲学论证，强调人类控制不仅要在形式上存在，更要在实质上有效——即人类确实拥有理解情境、作出判断并介入系统运行的能力与渠道[6]。可干预性是自主治理的“安全阀”：它保证了自主治理无论发展到何种程度，都始终处于人类可收回的范围内。在操作层面，可干预性可以通过最大干预延迟（如紧急中止的响应时限）与权限收回的机制完备性等指标加以检验。其中，权限收回的机制完备性要求在分布式系统中通过权限令牌与吊销机制实现授权撤销，从而确保人类的介入权在技术架构上始终可得。

责任可追溯性：损害能够追溯到责任主体。 当系统行为造成损害时，应当能够将损害追溯到明确的责任主体，而非消解于算法与数据的复杂性之中。Nissenbaum 从责任伦理角度指出，计算机系统所带来的一项根本挑战是“责任的消解”——责任在复杂的自动化链条中被稀释、转移乃至消失；而负责任的计算系统必须确保责任能够被追溯与归因[31]。Diakopoulos 进一步讨论了算法问责的实现机制，主张通过披露、审计、反馈与救济等途径将算法决策纳入可问责的框架[32]。在操作层面，责任可追溯性要求建立多主体协同决策下的责任归因仲裁机制，并明确损害追溯的时效与证据链要求，使归责不因系统与主体的复杂性而消解；这既是对“责任的消解”难题的制度性回应[31]，也是对分级授权中多主体协同场景的必要技术安排。

四个条件并非彼此独立：透明性提供理解基础，可审计性提供核查手段，可干预性提供纠错渠道，责任可追溯性提供归责终点。它们共同回答了“凭什么允许系统自主治理”这一正当性问题——只有当一个自主治理系统能够被理解、被审查、被介入、被追责时，社会才有理由赋予其治理层面的自主权。上述条件的操作化指标可以汇总为 Table 3。

Table 3 | 四个正当性条件的操作化指标与示例性阈值

正当性条件	操作化指标	测量方法	阈值设定与责任主体	示例性阈值（非强制性规范）
透明性	决策轨迹可导出粒度；可解释性强制下限；解释忠实性（理由与决策轨迹一致性）	高风险动作的轨迹导出测试；解释一致性抽检（含解释—轨迹一致性核验）	授权主体依据场景风险等级设定	重大影响动作 100% 可导出决策轨迹；95% 以上动作可给出自然语言行动理由；重大影响动作解释与轨迹一致率不低于 90%

正当性条件	操作化指标	测量方法	阈值设定与责任主体	示例性阈值（非强制性规范）
可审计性	日志保留周期；防篡改存储；独立审计频率与触发条件；审计主体独立性保障；治理有效性实测；审计结果公开披露	日志完整性校验；对抗性违规用例测试；审计机构利益冲突审查；审计报告公开核验	授权主体与独立审计机构共同设定	高风险日志保留不少于 5 年；授权升级前与重大事件后强制专项审计；审计机构与运营方无利益关联；审计结论公开披露
可干预性	最大干预延迟；权限收回机制完备性	紧急中止演练；权限吊销时效测试	授权主体设定	紧急中止延迟不超过 10 秒；权限吊销在限定时间内全链路生效
责任可追溯性	归因仲裁机制；追溯时效与证据链要求	多主体责任归因演练；证据链完整性核验	授权主体与监管主体共同设定	损害事件 100% 可归因至责任主体；证据链可回溯至原始日志

表注：本表为正当性条件的操作化度量提供参考框架。所列阈值为说明性的示例取值，仅用于展示“如何测量”，不构成强制性技术规范；实际阈值应由授权主体依据应用场景风险等级另行设定（见 5.5 节）。“测量方法”“阈值设定与责任主体”两列旨在落实可审计性与责任可追溯性的实现路径。

这些指标使四个条件从抽象原则落实为可检验、可审计、可验证的操作要求，也为下文分级授权中“治理闭环成熟度”的判断提供了度量基础。

5.3 行业自我规制的辅助定位

在正当性建构中，行业自我规制并非无关紧要，但它只能承担辅助功能。Mittelstadt 明确指出，“仅凭原则无法保障合乎伦理的人工智能”——伦理准则固然具有凝聚共识、指引设计的功能，但缺乏强制性实施机制，难以对系统的实际行为形成约束[33]。结合本文第 2 章与第 4 章的分析可以进一步定位行业自我规制的角色：它主要承担“设计阶段规范输入”——为系统的规则体系与价值设定提供规范性内容来源；以及“验证支撑”——为系统的治理闭环提供外部的检验、认证与问责参照。然而，行业自我规制无法替代系统内部的治理闭环：外部的规范若不能内化为系统的自我管理机制，就始终停留在倡导层面。因此，行业自我规制是正当性的必要补充，而非充分条件。

5.4 分级授权路径

正当性条件的满足是一个渐进过程，其实现路径在于分级授权。借鉴回应性监管（responsive regulation）的理论传统，Ayres 和 Braithwaite 主张监管强度应与被监管对象的自控能力相匹配——当主体展现出可靠的自律能力时，监管者可以给予更大的自主空间；反之则收紧控制[34]。这一逻辑映射到人工智能自主治理，从而形成三级授权路径：

在展开三级授权路径之前，有必要将本文的分级思路与主流分级框架作一对照[1]。

Table 4 | 本文分级授权与主流分级框架的对比

框架	分级依据	授权逻辑	动态性	与治理闭环的关系
欧盟《人工智能法案》	风险类别	风险越高，义务越重	以静态分类为主	提供横向风险分类
SAE J3016	自动化程度（人机角色分配）	自动化水平逐级递进	静态分级	关注执行自动化，不含治理维度
FDA AI/ML SaMD	风险类型与修改类型	变更需重新评估	部分动态	侧重上市后监督
本文三级授权	治理闭环成熟度	自证可治理则授权升级	动态授权；受禁止性边界约束	治理闭环即授权依据，正交于禁止性监管

与上述框架相比，本文分级路径的判据创新在于以“治理闭环成熟度”作为授权升级的依据：一个系统能否从“辅助决策”升格为“有限自主执行”乃至“高度自主治理”，不取决于其能力规模或所属风险标签，而取决于它能否通过可审计的证据证明自身已满足四个正当性条件。为使“成熟度”具备可度量的操作性，本文采用“环节最低门槛 + 条件综合评估”的评估框架：事前审核、事中监控应对、事后复盘提炼三个环节须分别达到既定的最低成熟度门槛，任一环节缺失则整体不达标；在门槛之上，再以四个正当性条件的操作化指标（见 Table 3）进行综合评估。成熟度的评估主体与评估程序由人类元治理的规则修订职能界定（与 Table 3 阈值设定主体一致），评估结论同时作为分级授权升级与降级决定的依据；各环节的具体权重与领域化校

[1] 欧盟《人工智能法案》依据风险将人工智能系统分为不可接受风险、高风险、有限风险与最小风险等类别，并据此施加差异化监管义务[7]；自动驾驶领域广泛采用的 SAE J3016 标准按人类与系统的角色分配将驾驶自动化分为 L0 至 L5 六级[35]；美国食品药品监督管理局则对人工智能/机器学习医疗器械提出基于风险与修改类型的监管路径[36]。三者的比较见 Table 4。

准属于后续经验研究的范畴，本文给出的是评估框架而非固定权重。具体而言，升级检验程序包括：前一授权等级下可审计行为记录完整且无重大违规；四个正当性条件的独立审计通过记录——审计须由与开发者及运营者无利益关联的第三方实施，并在启动前完成利益冲突审查；治理闭环各环节（事前、事中、事后）运作证据的持续积累，以及对闭环约束力的穿透式实测——即通过预设违规场景验证系统的监控与应对机制确能触发，而非仅在记录层面"看起来合规"；审计结论与授权依据应当公开披露，接受社会监督。这一判据使本文的分级路径具备动态性——授权不是一次性的标签，而是随治理证据的积累而演进。需要指出，本文分级与欧盟《人工智能法案》等横向监管框架并不冲突，而是互补：后者回答"哪些系统需要更严格的监管"，前者回答"一个系统如何在监管边界内逐步获得更多自主"。

第一级：辅助决策。 系统仅提供决策支持与建议，最终决策权保留于人类主体。这是自主治理化的起点，系统的治理功能主要体现为对自身建议过程的透明与可解释。

第二级：有限自主执行。 在明确限定且低风险的领域内，系统可以自主执行任务，但受制于严格的目标边界、规则约束与监控机制，关键节点仍需人类介入。治理闭环在这一级开始完整运转。

第三级：高度自主治理。 在积累充分验证与可信记录之后，系统可以在较高层次上自主管理自身行为，包括自主设定执行方案、自主监控与自主复盘。但即使在最高层级，其授权范围与边界仍由人类事先界定并保留监督。

分级授权的意义在于：它以风险与能力相匹配的方式逐步积累正当性——每一次授权升级都以系统在前一等级的可靠表现为依据，使自主治理的扩展建立在可验证的信任基础之上。

与授权升级相对称，分级授权还须包含降级与收回机制，否则授权轨迹将呈现"只进不退"的单向形态，这与回应性监管的核心逻辑——监管强度随被监管者自控表现的动态变化而双向调整——不相一致。授权降级或收回的触发条件包括四类：连续审计不通过（如连续两个评估周期未通过独立审计）、规范违反率超过既定阈值、治理有效性实测失败（穿透式实测未触发预期干预），以及重大责任事故或系统性规范偏离。任一条件触发，授权等级则应下调或暂停。降级决定由授权主体依据审计证据作出，须经独立审计复核并公开披露降级理由；当降级至最低等级仍不能恢复治理有效性时，则触发权限收回——此时与可干预性条件的技术保障（权限吊销）及人类元治理的紧急干预职能相衔接，形成"日常降级（制度性）→权限收回（技术性）→紧急干预（根本性）"的三级响应链。需要强调，降级与升级共用同一套治理证据体系：用于支持升级的可审计行为记录、独立审计通过记录、治理闭环运作证据与穿透式实测结果，同样构成降级决定的依据来源。由此，分级授权成为对称的双向动态机制——表现优良则积累正当性而升级，表现恶化则消耗正当性而降级。

需要进一步界定的是本文分级授权的适用边界。分级授权处理的是“在允许的自主范围内，系统可以逐步获得多少自主”这一问题，它预设了系统本身属于法律允许部署的对象；对于被禁止部署或使用的系统类型——例如欧盟《人工智能法案》所禁止的不可接受风险的人工智能实践——本文的分级授权路径不适用，系统也不会因自证可治理而获得合法地位。换言之，分级授权与禁止性监管正交：前者在合法空间内部依据治理闭环成熟度分配自主权，后者划定该空间的硬性边界。任何分级授权都不得突破禁止性边界，这一边界是下文所述人类元治理所应坚守的底线之一。

5.5 人类元治理的不可让渡

分级授权之上，还须确立一项不可让渡的底线——人类元治理。元治理（meta-governance）指对治理本身的治理，即设定治理框架、规则与边界的更高层级活动。Jessop 指出，元治理涉及对治理方式的反思性安排，包括对参与主体、规则结构与运行前提的组织与再组织[37]；Black 进一步表明，在多元、多中心的复杂治理体系中，元治理为协调不同治理主体、维护治理体系的一致性与合法性提供了关键机制[38]。将元治理理论引入人工智能自主治理，意味着以下三项职能必须保留在人类或社会层面，不可授权给系统：其一，价值设定——系统应当追求何种价值、遵循何种根本规范，只能由人类社会经由民主与伦理审议来确定，而不能由系统自我设定；其二，规则修订——治理闭环所依据的规则体系，其修改与废止权不能完全让渡给系统自身，否则系统将获得“自我立法”的能力，从而消解人类控制的意义；与规则修订相衔接，Table 3 所列正当性指标的具体阈值设定同样属于元治理范畴——由哪一主体依据何种风险等级设定何种阈值，属于对治理框架本身的安排，只能由人类或授权监管主体确定，而不能交由系统自我裁量；其三，紧急干预——在极端风险情形下，人类必须保有终止系统运行的最终权力，这一权力与可干预性条件相衔接，但层级更高、性质更根本。需要说明，人类元治理并非仅在极端时刻“在场”的应急机制，其运作具有日常的过程性：元治理既包括极端情形下的紧急干预，也包括对治理闭环运作的持续监督、对规则体系与正当性指标阈值的定期评估与调适——后者正是“规则修订”职能的常态化表现。这一过程性维度与可干预性条件形成层次衔接：可干预性保证人类“能够”在日常介入系统运行（提供介入渠道），元治理的规则修订则提供持续调适治理框架的规范权力；而在授权机制层面，制度性的日常降级、技术性的权限收回与根本性的紧急干预共同构成三级响应链（见 5.4 节），使人类控制既覆盖日常过程，又覆盖极端时刻。人类元治理的不可让渡性，为分级授权的进程划定了不可逾越的边界：自主治理可以逐步扩大，但治理的最终权威始终锚定于人类与社会。

上述讨论显性化了治理闭环与人类外部控制的关系：二者不是竞争关系，而是分层互补关系。系统内部的治理闭环位于操作层（operational level），负责对日常自主行动的管理；人类元治理位于元层（meta level），保留价值设定、规则修订与紧急干预三项职能；紧急干预正是连接两层的超越机制——当操作层的内部闭环失灵或出现极端风险时，元层自上而下接管，其行使不依赖系统的同意。这一分层结构同时界定了“系统拒绝修正”这一边界情形：当系统基于自身判断认为行动合规、

而人类认为需要中止或修正时，人类元治理的有效性恰恰体现在其干预不以系统认同为前提。换言之，可干预性条件（见 5.2 节）保证人类“能够”干预，人类元治理保证人类“有权”最终干预：前者是技术能力的保障，后者是规范权力的确立。自主治理无论如何发展，其治理闭环都处于元治理框架之内，这正是人类控制不可让渡的实质含义。上述分层结构可以示意为 Figure 2。

Figure 2 | 治理闭环与人类元治理的分层结构

Figure 2 | 治理闭环与人类元治理的分层结构：治理闭环（事前审核—事中监控应对—事后复盘提炼）位于操作层，人类元治理位于元层，保留价值设定、规则修订与紧急干预三项职能；紧急干预作为超越机制自上而下连接两层，人类控制不可让渡的边界以竖虚线标示，为治理闭环划定不可逾越的边界。

5.6 本文的规范边界

最后，有必要对本文的定位作如实说明。本文是一项规范性概念研究：其目标是为人工智能自主治理提供概念框架、正当性条件与实现路径的理论论证，而非对特定系统的经验评估或技术方案设计。因此，本文的治理闭环、四个正当性条件与分级授权路径应被理解为分析框架与规范主张，它们为经验研究与工程实现提供了概念坐标，但本身不构成可立即部署的技术规范。与上述定位相配套，本文框架的分析粒度需要显式说明：本文以“单个自主系统及其直接的人类元治理关系”为分析单元，治理闭环、四个正当性条件与分级授权均以单系统为对象。该框架适用于单系统作为独立治理单元的部署场景，可作为多智能体协作与系统生态治理的“单元层”基础；但对于系统生态的整体治理成熟度问题——包括系统间涌现行为与跨系统责任分配

——单系统框架不能单独回答，需待生态级治理理论的发展（相关展望见 4.4 节末）。本文第 4 章对自主编码智能体的示意性诊断展示了框架的诊断力，但系统级的实证检验——例如在真实部署环境中对治理闭环各环节的度量与审计——仍有待后续工作。明确这一边界有助于读者准确评估本文主张的适用范围，也避免了将规范性论证误读为经验结论。

---

6. 结论

本文沿着“人工智能—自主化—治理化”的三层递进关系，完成了对人工智能自主治理的概念建构。在这一链条中，每一层都不是对前一层的简单替代，而是能力与功能的叠加：人工智能提供了人类知识数字化基础上的统计推理能力，构成自主治理的知识底座；自主化在推理之上叠加执行能力，实现了从“生成结果”到“执行任务”的跨越；治理化则在执行之上叠加管理功能，形成了“事前审核—事中监控应对—事后复盘提炼”的治理闭环。

贯穿全文的一个核心区分是“执行自主”与“治理自主”。执行自主回答“怎么做”，是能力的扩展；治理自主回答“能不能做、做得对不对、出了问题怎么办”，是规范的嵌入。将治理化误认为更强大的执行能力，是人工智能治理讨论中最危险的混淆——它会使治理问题被错误地还原为技术问题，从而在能力竞赛中忽略了对规范机制的建设。本文强调：治理化不是技术能力的增强，而是管理功能的加入；自主治理化使人工智能从“受外部规制的执行者”转变为“能够自我管理的治理主体”。

自主治理化的正当性并非自动获得，它必须满足透明性、可审计性、可干预性、责任可追溯性四个条件。这四个条件共同回答了“凭什么允许系统自主治理”的正当性问题，也与“有意义的人类控制”的理论要求相互印证[6]。正当性的建立与扩展是一个渐进过程：通过“辅助决策—有限自主执行—高度自主治理”的分级授权路径，在风险与能力相匹配的原则下逐步积累可验证的信任记录；而人类元治理则构成了不可让渡的边界——价值设定、规则修订与紧急干预必须保留在人类或社会层面，使治理的最终权威始终锚定于人类社会[21], [37]。

就治理理论的整体图景而言，本文的框架亦是对既有伦理原则主张的一次推进。Floridi 和 Cowls 的五原则框架为人工智能治理确立了伦理层面的规范性基础，但原则要真正生效，必须内化为系统的运作机制[1]；本文所建构的治理闭环与正当性条件，正是这种内化的概念化表达。人工智能自主治理化的推进，应坚持“执行自主”与“治理自主”的区分，将治理化建立在技术、制度与规范三者耦合的基础上——技术提供自我管理的能力，制度提供授权与问责的结构，规范提供价值与方向的源泉。唯其如此，人工智能才能在日益自主的同时，始终处于可理解、可审查、可干预、可追责的治理秩序之中。

同时需要说明，本文作为一项规范性概念研究，其框架的价值在于为自主系统的设计、部署与监管提供概念坐标与论证依据；框架在真实系统上的经验检验——包括治理闭环各环节的度量方法与分级授权判据的实证验证——是本文后续研究的重要方向。

参考文献

[1] L. Floridi and J. Cowls, “A unified framework of five principles for AI in society,” Harvard Data Science Review, vol. 1, no. 1, 2019.

[2] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi, “Artificial intelligence and the ‘good society’: The US, EU, and UK approach,” Science and Engineering Ethics, vol. 24, no. 2, pp. 505–528, 2018.

[3] A. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” Nature Machine Intelligence, vol. 1, no. 9, pp. 389–399, 2019.

[4] T. Hagendorff, “The ethics of AI ethics: An evaluation of guidelines,” Minds and Machines, vol. 30, no. 1, pp. 99–120, 2020.

[5] Y. Bai et al., “Constitutional AI: Harmlessness from AI feedback,” arXiv:2212.08073, 2022. [Online]. Available: https://arxiv.org/abs/2212.08073

[6] F. Santoni de Sio and J. van den Hoven, “Meaningful human control over autonomous systems: A philosophical account,” Frontiers in Robotics and AI, vol. 5, art. 15, 2018.

[7] European Parliament and Council of the European Union, “Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act),” Official Journal of the European Union, 2024.

[8] OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449, 2019.

[9] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), National Institute of Standards and Technology, 2023.

[10] UNESCO, Recommendation on the Ethics of Artificial Intelligence, United Nations Educational, Scientific and Cultural Organization, 2021.

[11] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Pearson, 2020.

[12] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[13] A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems, vol. 30, 2017.

[14] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in Proc. FAccT ’21, 2021.

[15] J. Kaplan et al., “Scaling laws for neural language models,” arXiv:2001.08361, 2020. [Online]. Available: https://arxiv.org/abs/2001.08361

[16] M. Wooldridge and N. R. Jennings, “Intelligent agents: Theory and practice,” The Knowledge Engineering Review, vol. 10, no. 2, pp. 115–152, 1995.

[17] E. Karpas et al., “MRKL systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning,” arXiv:2205.00445, 2022. [Online]. Available: https://arxiv.org/abs/2205.00445

[18] T. Schick et al., “Toolformer: Language models can teach themselves to use tools,” arXiv:2302.04761, 2023. [Online]. Available: https://arxiv.org/abs/2302.04761

[19] S. Yao et al., “ReAct: Synergizing reasoning and acting in language models,” in Proc. Int. Conf. Learning Representations (ICLR), 2023.

[20] S. Russell, Human Compatible: Artificial Intelligence and the Problem of Control. Viking, 2019.

[21] E. Ostrom, Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, 1990.

[22] I. Rahwan, “Society-in-the-loop: Programming the algorithmic social contract,” Ethics and Information Technology, vol. 20, no. 1, pp. 5–14, 2018.

[23] J. García and F. Fernández, “A comprehensive survey on safe reinforcement learning,” Journal of Machine Learning Research, vol. 16, no. 1, pp. 1437–1480, 2015.

[24] M. Alshiekh et al., “Safe reinforcement learning via shielding,” in Proc. AAAI Conf. Artificial Intelligence, 2018.

[25] P. F. Christiano et al., “Deep reinforcement learning from human preferences,” in Advances in Neural Information Processing Systems, vol. 30, 2017.

[26] M. Chen et al., “Evaluating large language models trained on code,” arXiv:2107.03374, 2021. [Online]. Available: https://arxiv.org/abs/2107.03374

[27] C. E. Jimenez et al., “SWE-bench: Can language models resolve real-world GitHub issues?” in Proc. Int. Conf. Learning Representations (ICLR), 2024.

[28] D. Amodei et al., “Concrete problems in AI safety,” arXiv:1606.06565, 2016. [Online]. Available: https://arxiv.org/abs/1606.06565

[29] J. Zerilli, A. Knott, J. Maclaurin, and C. Gavaghan, “Transparency in algorithmic and human decision-making: Is there a double standard?” Philosophy & Technology, vol. 32, no. 4, pp. 661–683, 2019.

[30] J. A. Kroll et al., “Accountable algorithms,” University of Pennsylvania Law Review, vol. 165, pp. 633–705, 2017.

[31] H. Nissenbaum, “Accountability in a computerized society,” Science and Engineering Ethics, vol. 2, no. 1, pp. 25–42, 1996.

[32] N. Diakopoulos, “Accountability in algorithmic decision making,” Communications of the ACM, vol. 59, no. 2, pp. 56–62, 2016.

[33] B. Mittelstadt, “Principles alone cannot guarantee ethical AI,” Nature Machine Intelligence, vol. 1, no. 11, pp. 501–507, 2019.

[34] I. Ayres and J. Braithwaite, Responsive Regulation: Transcending the Deregulation Debate. Oxford University Press, 1992.

[35] SAE International, Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (SAE J3016). SAE International, 2021.

[36] U.S. Food and Drug Administration, Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD). U.S. FDA, 2021.

[37] B. Jessop, “Governance and meta-governance: On reflexivity, requisite variety, and requisite irony,” in Governance as Social and Political Communication, H. P. Bang, Ed. Manchester University Press, 2003.

[38] J. Black, “Constructing and contesting legitimacy and accountability in polycentric regulatory regimes,” Regulation & Governance, vol. 2, no. 2, pp. 137–164, 2008.