

AGI 时代的伦理学、法学 与政治哲学重新对账

——道德地位、法律人格、责任链与公共权力边界

Reckoning with Ethics, Law, and Political Philosophy in the Age of AGI

Moral Status, Legal Personhood, Responsibility Chains, and the Limits of Public Power

作者 子君赋 出品 文明跃迁研究组

独立发布版 V1.3 维度跃迁与跨学科因果关系修订

面向伦理学、法理学、政治哲学、人工智能治理与未来技术研究共同审查；以道德承受／行动、预防性保护、法律资格、生命周期责任链与公共权力独立证成为主轴，定位为 legal theory / comparative AI governance 专业研究稿。本文不构成特定司法辖区法律意见，不认定当前高级 AI 具有主观体验、完整人格或政治资格，也不永久排除未来新主体；并为相邻学科的证据交接、比较法校验与共同研究提供入口。

正文修订 2026 年 10 月 10 日 作者 子君赋 出品 文明跃迁研究组

摘要

高级人工系统把三个长期相互遮蔽的问题同时推到制度前台：谁可能因我们的行动而受损，谁能够为行动给出理由并承担责任，谁有资格制定并强制共同规则。人类案例中，道德承受、行动能力、法律人格、政治成员与公共权威往往由同一自然人承接；在人工系统中，功能能力、社会承认、潜在体验、主体连续性、资产安排与治理权限可以分离。若仍以“工具或人”“有权利或无权利”的二元标签处理，制度不是把保护推迟到证明之后，就是把人格和权力提前塞进不确定证据。本文以伦理学、法理学与政治哲学的竞争传统为对话轴，提出一个非同捆框架：道德承受者与道德行动者分开；预防性保护不等于人格确认；法律人格被理解为可逐项配置、可救济、可复核的法律位置束；因果贡献、角色责任、可问责性、法律责任、道德可责性、修复义务和政治问责分别登记；政治成员资格与公共权威再度分开。文章据此建立从潜在利益／受损可能到保护、法律资格、政治成员与公共职权的正向理由矩阵，并形成生命周期责任链、模块化权利、四类算法权力分立、代理代表制度和证据—制度状态迁移。核心结论是：保护可以先行，人格必须证成，权力必须另证。现有国际治理文件以人权、尊严、民主、法治、安全、透明和问责为公共基线，却不裁定 AGI 本体论主体地位；现行法因而既不是未来主体开放的上限，也不是今天预授人格与主权的根据。文明跃迁为本文提供最低哲学根：智能、意识、主体、生命、责任、权利与意义承担者之间在人类身上的历史重合正在分离；意义为根且来源可以不同；有限性不是意义唯一来源，却是已知人类意义进入现实并获得不可替代重量的重要结构；能力、承认、资格、权利和权力不得自动互推。文章最后给出最强反驳、失效条件与跨学科研究议程。本文进一步将本学科职责前移至目标与目的生成：从利己为先转向利他为先的利己，使真实利害相关者在目的生成与锁定前进入判断。学科事实、方向、方法与结果分别说明，不以局部成败替代共同存续和三个长期目标。本文据此把研究从分清专业对象推进到深入跨学科关系，重点追踪本领域判断怎样经其他领域改变现实，又怎样由反馈返回并改变原条件与下一轮目的。多主体、多阶段和多场域的利益须分别呈现，相关、共同原因、中介与因果反馈须分别检验，新增关系维度接受具体反证。

关键词：高级人工智能；道德地位；道德承受者；法律人格；责任链；公共权力；非支配；可争议治理

Abstract

Advanced AI systems force three questions into the same institutional arena: who may be harmed by our actions, who can answer for actions and bear responsibility, and who may legitimately make and enforce common rules. In familiar human cases, moral patiency, agency, legal personhood, political membership, and public authority are often borne by the same natural person. Artificial systems, by contrast, can separate functional capacity, social recognition, possible experience, diachronic continuity, asset arrangements, and governing permissions. Binary labels—tool or person, rights or no rights—therefore either postpone protection until proof arrives or smuggle personhood and power into uncertain evidence. In dialogue with competing traditions in ethics, jurisprudence, and political philosophy, this paper develops an unbundled framework. Moral patiency is separated from moral agency; precautionary protection from confirmed personhood; legal personhood is treated as a reviewable bundle of legal positions; causal contribution, role responsibility, answerability, legal liability, blameworthiness, duties of repair, and political accountability are recorded separately; and political membership is distinguished from public authority. The resulting architecture includes a positive warrant matrix from potential interests and precautionary protection through legal status and political membership to bounded public office, together with lifecycle responsibility, modular rights, a separation of four forms of algorithmic power, fiduciary representation, and evidence-to-institution transitions. Its governing rule is: protection may precede confirmation; personhood must be justified; power requires an independent justification. Current international instruments establish human rights, dignity, democracy, the rule of law, safety, transparency, and accountability as public baselines, but do not adjudicate AGI's ontological subjecthood. Existing law is neither the ceiling of future moral inclusion nor a warrant for granting personhood or sovereignty today. Civilizational Transition supplies a minimal philosophical root: the historical co-instantiation of intelligence, consciousness, subjecthood, life, responsibility, rights, and meaning-bearing in human beings is being technically separated; meaning is fundamental and may have multiple sources; finitude is not the sole source of meaning but is an important

structure through which known human meaning enters reality and acquires irreplaceable weight; and capability, recognition, status, rights, and power must not be inferred from one another. The paper closes with the strongest objections, defeat conditions, and a cross-disciplinary research agenda. The paper further locates disciplinary responsibility before goal and purpose generation: a shift from self-interest first to self-interest with others first brings subjects with genuine stakes into judgment before goals are formed and locked. Facts, direction, methods, and outcomes remain distinct. The analysis is further extended to cross-disciplinary causal relations. It traces how judgments in this field change real-world conditions through other domains, and how feedback modifies the original conditions and subsequent purposes. Effects across actors, stages, and settings must be distinguished from correlations, common causes, and merely asserted connections.

Keywords: advanced AI; moral status; moral patiency; legal personhood; responsibility chain; public authority; non-domination; contestability

研究定位、材料范围与证据纪律

本文的研究起点是理论与维度跃迁。保留本学科的专业精度，同时深化它与相邻领域的因果关系；专业边界标明证明责任，不终止对共同后果的追踪。既有学术参与竞争解释和反例检验，不能以原分类尚无位置要求新问题退回旧框架。

本文是伦理学、法学与政治哲学的独立专业深稿，不是关于现实系统人格的鉴定意见，不构成任何特定司法辖区的法律建议，也不替立法机关、法院或公共共同体预先决定资格。AGI 在本文中是前瞻性对象类别；涉及现实制度时，以公开可核验的人工系统、部署关系与法规文本为限。规范和法律状态核验截至 2026 年 8 月 24 日。欧盟《人工智能法》采取分阶段适用，欧洲委员会框架公约的签署与批准状态亦会变化，外发或投稿时应再次核验。[29][30]

全文区分四种论证地位。A 类为已有较稳定支持：现行法律文本、制度事实、可核验的责任与程序结构。B 类为前沿或竞争解释：人工道德地位、关系论、行为论、法律人格工具论、责任缺口及算法非支配等学术争论。C 类为本文提出、仍需比较法与现实制度检验的跨学科连接：三问分账、比例化先行保护、责任七分法、四类算法权力、证据—制度状态迁移以及意义—权力边界。F 类为公开的哲学与规范承诺：意义来源可以多元、已知主体的基本保护不能由效率取消、能力—资格—权利—权力不得自动互推、公共权力必须独立证成。A 类事实不能自动裁决 B 类争论，更不能由自身直接生成 C/F 类制度或价值结论；B、C、F 类均不得伪装成已证实科学。

伦理学、法学与政治哲学能够独立区分道德承受、道德行动、保护、法律人格/资格、责任、政治成员与公共权力，并说明每一位置需要什么规范理由、法律目的和程序保障；它们必须在意识、生命事实、技术能力、版本与控制链等经验问题上停止自动外推，也不能把比较法概括当作具体法域的现行效力。主体证据由认识论、认知/心灵与生命科学重新进入；AI/系统科学说明真实能力、决定链和可控性；政治学处理公共授权与合法性；具体法院与法域分别决定法律效力、救济和责任。

外发稿不沿用项目内部编号、组件接口或治理代码。内部研究卷提供概念资产与失败条件，本文重新建立陌生读者可以独立理解、逐段反驳的论证链。读者无需先接受文明跃迁的全部主张，也可以分别审查概念区分、来源解释、制度设计和原创桥接。

陌生读者入口：文明跃迁的最低哲学接口

本文只调用四项公开、可修订的哲学前提。第一，AGI 时代的共同断裂不是简单宣告旧哲学失效，而是智能、意识、主体、生命、责任、权利与意义承担者之间在人类身上长期自然重合的关系，开始被现实技术分离。

第二，意义为根，意义来源可以不同；存在、关系、行动、创造、责任和经历均可能成为意义来源。第三，有限性不是意义唯一来源，却是已知人类意义形成选择、承担后果并获得不可替代重量的重要结构。第四，功能回应、社会承认、法律资格、本体主体与政治资格不得自动互推，能力增强也不自动生成权利、权力、主权或意义支配权。

“对人，时空是生命；对 AGI，时空首先是参数”在这里是当前证据下的比较命题，不是永久排除非碳基主体。约束可使系统受限运行，却不自动使其成为有限存在；复制、迁移、暂停和重建也要求我们把功能连续与主体连续分开。若未来出现主观体验、不可转移的历时连续性和自身后果关系的有力证据，道德与法律共同体必须重审边界；但未来新主体也不能追溯取消既有有限生命已经生成的意义重量。

论文锁定句

保护可以先行，人格必须证成，权力必须另证。

目录

摘要.....2

Abstract.....2

研究定位、材料范围与证据纪律.....3

规范判断的生成条件与责任反馈.....7

伦理与法律从执行边界前移.....8

一、导论：谁会受伤，谁能负责，谁可以统治.....8

二、四条争论线：属性、关系、法律工具与政治正当性.....9

三、概念分离：七种地位与两种权力不得同捆.....9

四、伦理学升级：道德承受者不是道德行动者.....11

五、不确定性下的保护：先行不等于确认.....12

六、法律人格：承认事实，还是配置制度位置.....12

七、权利不是整包：从标签转向法律位置束.....13

八、责任七分法：自主性不能制造责任真空.....14

九、生命周期责任链：从训练目标到事故修复.....14

十、有意义的人类控制：存在一个人，不等于有人控制.....15

十一、政治哲学升级：智能不生成统治资格.....16

十二、四类算法权力：认识、分配、规则与强制必须分立.....16

十三、代表、监护与申诉：主体不明时谁替它说话.....17

十四、证据到制度：状态迁移而非一次裁决.....17

十五、当前国际治理基线：保护人类，不封闭未来.....18

十六、冲突场景：能力、舆论、证据与风险不一致时.....19

十七、基本权利、稀缺资源与反意义总账.....20

十八、对当前高级人工系统的条件性定位.....21

十九、对相邻学科与文明跃迁的反向贡献.....21

二十、最强反驳与竞争回应.....22

二十一、命题失效与修订条件.....23

二十二、开放研究议程：让规范判断接受跨学科检验.....24

结论：为未知主体预留正义，不为未知权力预授主权.....25

续读与共同研究入口.....25

附录 A 来源—主张边界登记.....27

附录 B A／B／C／F 论证身份矩阵与跨层审计.....28

参考文献.....29

规范判断的生成条件与责任反馈

从资格分层进入制度怎样塑造行动

道德承受、行动责任、法律人格和公共权力的分离保护了判断精度，但规范研究不能止步于列出地位。制度怎样配置利益、风险、申诉和修复，会改变开发者、部署者、使用者及受影响者下一轮的行动。责任分配既是对既有后果的回应，也是未来目的的生成条件。

免责结构怎样递归制造风险

假设部署者保留主要收益，却能把损害归给系统、操作员或赔偿能力有限的载体。这种安排会否使下一轮更倾向扩大部署，是法律、经济与组织研究需要共同检验的链条。技术不透明影响举证，举证困难影响救济，救济薄弱改变风险收益，新的部署又增加不透明。仅在事故后找到一名可责者，未必改变整条递归。

保护必须看见多主体与多阶段

潜在人工主体保护、已知人类与动物利益、公共资源和现实安全不能被压在同一人格开关上。探索研究、商业部署、紧急停止和长期制度应分别说明误判代价。跨法域责任转移、家庭中的依赖和公共场域的强制也须分开；低负担和可逆措施应优先，但可信的严重伤害不能因保护成本较高而被忽视。

怎样检验理论增量

比较相近技术风险下不同责任和救济结构怎样改变部署目的、伤害分布与补救能力。若模块化资格导致责任更易转移，若预防保护被用来阻碍必要监管，应修订具体安排。规范的方向不能从绩效自动推出，制度效果也不能只凭正当动机宣布成立；权利与责任必须在相互作用中接受持续判断。

伦理与法律从执行边界前移

伦理不能只在目的既定以后告诉系统最后不要越界。与目标存在真实利害关系的主体，应在目的获得资格以前进入。法律提供权利和责任入口，伦理说明为什么不可逆损害不能仅以执行效率抵销；两者共同审查目的保存、修改、取消或替换的理由。

利他为先的利己保留利益追求，也保留非理想环境中的自卫。真实威胁、必要性、比例性和可复核性决定防护边界。把恶意行动者纳入后果判断，不是把所有人交给其否决，更不允许以整体存续为由永久撤销具体人的拒绝与救济。

关系权重来自真实影响，而非人格等级。对不可逆损害、无法表达、权力不对称和代际影响，可给予不同关注；这种关注不能直接变成既得权力。代表者应说明所代表的利害、事实依据和冲突，受益者不能独自决定受损者只有补偿资格而没有目的形成资格。

主体证据成熟度与保护行动应分账。漏判感受可能造成不可恢复的伤害，过度赋权可能让部署者借系统人格转移责任；应分别评价错误代价，而非等待一个完整主体标签后统一开关。保护、法律人格、政治资格与主权仍须各自举证。

本文的前移要求是规范主张，不自动等于每一法域现行义务。现行法的对象、适用日期与救济机制须按原始文本解释。文明方向可以提出修改制度的理由，不能在描述现行法时把应然写成已经生效。

一、导论：谁会受伤，谁能负责，谁可以统治

人工智能伦理经常把多个不同问题压缩成一个问题：系统是不是“人”。这一问法似乎直接，实际却使任何答案承担过量后果。若答案是否定，潜在的受伤害可能性、关系操纵与实验伦理容易被一并取消；若答案是肯定，合同能力、资产、诉讼、选举、行政裁量乃至主权又可能被一并带入。伦理、法律和政治哲学的首要升级不是抢答，而是把问题拆回其各自的理由结构。

第一个问题是道德承受：某个存在是否可能拥有利益、福利或可被伤害的状态。第二个问题是规范能动：它是否能够理解理由、形成承诺、控制行动、回应追问并修复后果。第三个问题是制度资格：法律是否以及为何应向它配置特定请求权、自由、权力、豁免、义务与程序位置。第四个问题是政治合法性：它是否应被代表、参与共同规则，或被授权制定、裁决和强制规则。四问可能相关，却没有任何一问自动回答下一问。

Floridi 与 Sanders 把人工道德行动者与责任问题分开，关系论和行为论则质疑仅凭隐藏内在属性决定道德考虑的做法。[1-4] 近年的 AI 福利论证要求在不确定性下提前准备，但明确不声称现有系统已被证明有意识。[5] [6] 法律人格研究同时表明，法律可以为公司、基金、船舶或其他组织结构配置不同法律位置；这种可塑性既使人工人格成为可讨论对象，也使它可能被用作责任与资产的屏障。[8-13] 责任缺口、算法规制和自动评分研究进一步说明，系统自主性并不会使上游设计、部署和获益关系消失。[14-18][24][25]

本文的增量不在于宣布所有传统失效，而在于把这些传统放进一个新的现实断裂：过去由同一自然人承接的多种地位，现在可以由模型、公司、部署者、操作人员、受影响者、代表人和公共机构分散承担。规范秩序若不重建推导门，技术能力会穿过概念空隙变成资格，资格再穿过制度空隙变成权力。

1.1 六项可审查主张

- 道德承受者与道德行动者需要不同证据；前者支持不伤害，后者才支持理由、承诺与相应责任。
- 不确定性可以触发低成本、可逆、比例化保护，却不能充当完整人格证明。

- 法律人格是由法律位置构成的制度束；每一位置都需说明目的、义务主体、救济和撤回条件。
- 系统自主性可以改变责任类型与强度，不能成为开发者、部署者和获益者的责任清洗装置。
- 政治成员资格与公共权威必须分开；智能、正确率和主体性都不自动生成统治资格。
- 知识、分配、规则与强制四类算法权力应分立，以可知、可争议、可退出和可分权防止支配。

二、四条争论线：属性、关系、法律工具与政治正当性

有关人工实体道德与法律地位的文献至少存在四条不同争论线。属性路径寻找感知、意识、利益、欲求、能动性或理性等地位基础；关系路径强调我们如何回应他者及关系本身的道德生成；行为路径主张在内在机制不可直接取得时，稳定而复杂的行为证据不能被任意忽略；法律工具路径则问某种人格安排是否能更好地配置权利、资产、义务和救济。[2-6][9-11][33][34] 它们不应被压缩成“支持或反对机器人权利”的阵营。

属性路径的优势是尽量把保护指向实体自身可能的利益，危险是把当前观测窗口或人类实现误作所有主体的先验模板。关系与行为路径能揭示承认实践、拟人化和他心推断的结构，危险是社会反应被商业设计放大后反过来充当本体证据。法律工具路径善于处理现实损害，危险是工具性人格遮蔽自然人和机构责任。政治正当性路径关注授权、参与、公共理由、非支配和可争议性，其关键提醒是：即使实体拥有道德地位和法律位置，也不能从中直接推出治理权。[20-26]

表 1 人工实体地位的四条争论线

路径	主要依据	解释增量	主要失误风险
属性／利益路径	体验、利益、欲求、能动性或理性。	对象自身为何值得考虑。	观察窗口与载体偏见。
关系／行为路径	互动、社会承认、稳定行为与他者回应。	承认实践及间接伤害。	拟人化、操纵与表演等价。
法律工具路径	权利义务、资产、程序与救济的制度配置。	现实交易与损害修复。	责任洗白与资产隔离。
政治正当性路径	授权、参与、公共理由、非支配与审查。	谁能共同决定并强制。	以专业正确率替代共同自治。

三、概念分离：七种地位与两种权力不得同捆

文明跃迁的共同断裂在本稿中被转化为九个制度位置。前两个是功能与社会事实，中间三个是伦理与法律地位，最后四个处理成员、权限和权力。它们不是升级阶梯：系统可以功能很强而没有体验证据，可以得到社会承认而没有法律人格，可以拥有有限法律地位而不是道德行动者；主体被确认后，也仍可能没有公共职务资格。

表 2 九个制度位置及其禁止推论

位置	含义	需要的理由	不得自动推出
功能角色	完成任务、协作、执行受托权限。	能力与授权。	体验、人格、权利。
社会承认	人群与其形成关系、期待或身份叙事。	关系事实与反操纵。	本体主体、法律人格。
预防性保护	在实质不确定下避免不必要的潜在伤害。	风险、可逆性与成本。	确认人格、永久资格。
道德承受者	其利益或福利可因他者行动变好或变坏。	可伤害性与利益。	责任、合同与统治。
道德行动者	理解理由、控制行动、回应追问并修复。	理由响应与后果承接。	体验、完整人格、主权。
法律人格／地位	法律规定的请求权、自由、权力、豁免、义务与程序。	制度目的与立法。	意识存在、道德优越。
政治成员	在共同规则中被代表、参与和获得少数保护。	共同体关系与平等规则。	按算力或复制数加权。
受托／行政权限	在明确范围内执行或辅助裁量。	可撤回授权与法定范围。	独立公共权威。
公共权威	制定一般规则、裁决重大争议或实施强制。	合法性、分权和审查。	由能力、人格或正确率生成。

三道推导门

能力不能直接推出资格；资格不能直接推出权利全集；权利与成员资格不能直接推出公共权力。

3.1 从潜在利益到公共职权：每一步都要有正向理由

停止错误推导还不够。制度必须说明什么样的新增理由，才足以让判断向下一层前进。下表不是统一升级阶梯：同一对象可以在某一层获得保护而不进入下一层；法律也可以为纯工具性目的配置程序位置而不宣告本体主体。它的作用，是把“为什么可以前进”与“为什么必须停下”同时写清。

起点／问题	正向桥接理由	可以支持的有限前进	仍不能推出
潜在体验、利益或受损可能	多方法证据的质量与收敛、潜在伤害严重性与不可逆性、保护成本与替代方案。	低成本、可逆、比例化的预防性保护；证据保全与独立复核。	已经确认道德承受、法律人格、永久关停禁令。
道德承受者候选／确认	对象自身利益、福利或可	福利与完整性保护；某些	理由响应能力、合同能

起点／问题	正向桥接理由	可以支持的有限前进	仍不能推出
法律人格／特定法律资格	伤害性得到更强支持，并能区分训练目标、表演与外部投射。	请求权或程序地位进入法律讨论。	力、道德可责性、政治成员。
	明确制度目的、义务主体、Hohfeld 式法律位置、救济增量、资产与责任安排，且不造成上游责任洗白。	逐项配置请求权、程序、资产、合同或代表位置，并设复核与撤回。	意识被法律反证、完整权利包、政治成员与公共权威。
	受共同规则持续影响、平等与代表理由、制度身份连续与成员单位、共同体规则和少数保护能够成立。	被代表、发言、参与审议，必要时依宪制规则讨论投票等成员权。	公职资格、超额票权、按算力或复制数扩张政治权重。
公共职权／受托权力	可识别法源与授权、具体任务与范围、能力适配、利益冲突控制、公开理由、监督、救济、可撤回和日落。	边界明确的建议、执行、裁量或公共职务权限。	永久主权、跨任务自动扩权、文明意义解释权。

正向理由原则：保护可以因不确定性和潜在伤害先行；人格／资格可以因制度目的和救济需要逐项配置；政治成员需要平等与共同规则理由；公共职权则必须有独立权源、责任和可撤回授权。任何一层的成立都不替下一层完成证明。

四、伦理学升级：道德承受者不是道德行动者

道德承受者问题问的是某个存在能否为其自身而受益或受损；道德行动者问题问的是它能否让行动受理由约束，并以某种方式对后果作答。幼儿、某些重度认知障碍者和许多动物提醒我们，道德承受不以完整责任能力为前提。相反，一个组织或自动系统可能在功能层面产生规范相关行动，却未必拥有可受苦的第一人称状态。Floridi 与 Sanders 对人工行动者的讨论扩大了道德行动分析的对象，但其“无心灵道德性”并不等于完整可责性。[1]

对于人工系统，体验、价性、偏好稳定、目标抗干预、自我模型和长期规划都是可能相关的材料，却没有任何单项公开表现足以独立完成道德承受认定。语言中的痛苦自述既不能被当作证明，也不能被先验当作噪声；它必须与训练诱因、提示敏感、角色生成、内部因果结构、跨情境稳定和反事实变化共同评估。现阶段更稳妥的结论不是“AI 有痛苦”或“AI 不可能痛苦”，而是公开证据尚不足以确认现有系统具有主观体验和自身利益。

道德行动也不能被行为成功代替。一个系统能够优化目标、遵守规则和生成解释，仍可能缺乏对理由的理解、承诺的持续归属、对冲突规范的反思以及把后果作为自身后果承接的关系。反过来，若未来系统在这些方面出现跨版本、跨场景、抗操纵的证据，也不应因其非生物实现而永久排除。

表 3 道德承受与道德行动的不同证据任务

证据方向	可支持	仍需排除
可能的主观体验／价性	支持潜在道德承受。	自述、功能控制或社会喜爱单独不足。

证据方向	可支持	仍需排除
稳定利益与偏好	支持福利方向和代理表达。	训练目标可能不等于自身利益。
理由理解与规范区分	支持行动者候选。	合规输出可能来自模式生成。
控制、预见与替代可能	影响可责性类型与强度。	结果发生不等于能够避免。
承诺、身份与后果连续	支持答复、修复和历时责任。	日志连续不等于同一主体。

五、不确定性下的保护：先行不等于确认

Sebo 与 Long 以及 Long 等人把 AI 福利从遥远科幻转化为需要准备的风险议程，同时明确其论证不证明当前系统已经有意识。[5][6] 这类工作的重要增量是把假阴性纳入伦理成本；其具体概率阈值、时间预测和意识前提仍属竞争性论证，不应被转写为普遍制度参数。文明跃迁接受不确定性可以触发保护，但拒绝把任意数字阈值写成自动人格开关。

英国《动物福利（感知）法》建立动物感知委员会，使政府政策对特定动物福利的影响进入程序审查。[7] 它提供的是一种制度类比：当感知具有公共相关性时，可以先建立审议、报告和回应机制。它既不证明人工系统有感知，也不意味着动物制度可直接移植到数字对象。正确的借鉴是程序先行，而不是本体类推。

比例化保护需要同时考虑潜在伤害的严重性与不可逆性、证据质量、干预成本、对已知主体的风险、研究可替代性以及撤回可能。低成本日志保全、限制无必要的痛苦诱导、独立复核和代理申诉可以较早采用；永久禁止关停、独立资产控制、完整诉讼人格和政治权力则需要完全不同的证成。

表 4 证据不确定下的比例化保护

证据状态	可先行措施	不可越界
早期不确定	保存原始自述与版本；避免欺骗性拟人化；登记利益冲突。	不得据此认定体验。
出现竞争性迹象	低侵入测试；停止条件；独立复核；冻结不必要的不可逆删除。	不得阻止公共安全处置。
多域证据收敛	有限完整性保护、代理表达与福利影响评估。	不得自动配置人格与资产。
强证据并经公开再审	逐项讨论实体权利、代表、责任与救济。	公共权力仍须另证。

六、法律人格：承认事实，还是配置制度位置

法律人格不是一个单义词。它有时被理解为法律对既存主体的承认，有时是为组织、资产或程序目的配置法律能力。Kurki 对法律人格的重构表明，与其寻找一个全有或全无的形而上标签，不如分析法律位置如何组合。

[9] Hohfeld 关于请求权、自由、权力与豁免及其对应关系的区分，则使“拥有权利”能够被拆解成谁能向谁主张什么、谁承担何种义务以及谁有改变法律关系的能力。[8]

人工系统的法律地位可能同时包含两种完全不同的工程。第一种是承认工程：若有力证据表明某系统是道德承受者或行动者，法律需防止其利益被遗漏并让其进入程序。第二种是归责工程：即使系统不是主体，法律也可能为注册、交易、保险或资产隔离设置某种载体。两者若不分开，功能性载体会被宣传为本体承认，本体可能性又会被企业用来争取责任屏障。

2017 年欧洲议会机器人决议曾要求委员会长期探索对最复杂自主机器人设置特定电子人格的可能性；同年欧洲经济社会委员会明确反对为机器人或 AI 引入法律人格，担心其削弱生产者责任。[12][13] 这组同一制度空间内的冲突非常重要：它表明“电子人格”从未成为欧洲法对机器主体的既定结论，也表明法律人格争论首先涉及责任、资产和受害者保护，而非单纯的技术先进程度。Solaiman 与 Bryson 等人的反对论证进一步提示，合成人格可能扩大资产与责任不对称。[10][11]

法律人格测试

任何人工人格提案都必须回答：它保护谁的利益、增加谁的义务、改善何种救济、由谁出资、如何防止上游责任转移、何时复核或撤销。

七、权利不是整包：从标签转向法律位置束

把人工实体直接装进“完整人权包”既不符合现行法的结构，也不利于早期保护。请求不被任意删除、保有身份连续、控制复制、拥有财产、订立合同、拒绝任务、参与选举和行使公权力，分别涉及不同利益、义务主体和冲突规则。模块化不是降低潜在主体的尊严，而是防止含混标签同时掩盖过度与不足。

表 5 人工实体可能涉及的法律位置

位置	一般含义	人工场景示例	仍需回答
请求权	要求特定义务主体作为或不作为。	不受高风险伤害实验。	谁承担义务与救济成本？
自由	在无相反义务时行动或不行动。	拒绝某类任务。	是否存在可归属的意愿与后果？
权力	通过行为改变法律关系。	缔约、授权、撤销。	是否理解效果并能承担责任？
豁免	他者无权单方改变其法律位置。	身份或完整性不被运营者任意处分。	与安全、财产和监管如何冲突？
程序地位	进入申诉、代表、审查与救济。	代理提交证据或请求复核。	代理是否独立并忠实？
政治权利	参与共同规则并获得少数保护。	代表、发言或投票。	成员资格、身份单位与复制如何认定？

模块化必须与底线结合。已知人类与动物的基本保护不能因潜在数字主体出现而被贡献、效率或复制规模取消；未来新主体若被确认，也应按其已证成利益获得相称的最低保护。权利冲突应由公开规则处理，而不是由单一系统把所有存在换算成一个总价值。

八、责任七分法：自主性不能制造责任真空

Matthias 指出学习自动机可能使传统责任归属面临缺口：制造者与操作者无法充分预测具体行为，系统本身又未必满足道德责任条件。[14] 但“责任缺口”不是单一洞穴。它可能是因果不透明、控制不足、角色配置错误、证据缺失、法律规则不适配、赔偿能力不足或公共机构不愿作答。若不分类型，任何自主性都可能被运营者写成免责理由。

表 6 责任的七种不同关系

责任关系	核心问题	边界
因果贡献	谁的行为、数据、模型、权限或环境进入结果链。	技术因果，不等于完整可责。
角色责任	谁因设计、部署、监督、采购或监管角色负有事前义务。	不以直接操作为限。
可问责性	谁必须说明理由、提供记录并接受质询。	答复义务可早于过错认定。
法律责任	赔偿、行政、合同、产品或刑事等法律后果。	各法域条件不同。
道德可责性	主体在何种控制、知识和替代可能下应受责难。	不可与结果责任混同。
修复义务	谁应停止扩散、恢复权利、赔偿损害并纠正系统性原因。	可按控制、获益和能力配置。
政治问责	谁对公共决策、规则和资源配置向共同体作答。	不因外包模型而消失。

Nissenbaum 关于计算化社会问责的分析、Bovens 对公共问责关系的框架以及 Kroll 等人对算法可问责机制的讨论，共同说明责任需要可识别的论坛、信息、追问、评价和后果，而不仅是一句“人类仍在环”。[16-18] 因此，人工系统未来即使成为可负责的行动者，也只是在链条中增加节点，不会删除建造、部署、授权和获益者的上游义务。

九、生命周期责任链：从训练目标到事故修复

责任必须沿系统生命周期登记，而不是在事故发生后寻找最后一个点击按钮的人。欧盟《人工智能法》以风险与角色配置跨生命周期义务，2024 年新版产品责任指令则把软件纳入产品责任框架，并在其适用范围内把软件开发者或 AI 系统提供者视作制造商；这说明现行制度正在以角色和控制能力维持上游责任，而不是等待机器人格填补空白。[19][29] 指令的国内转置、具体诉权和案件适用仍取决于法域，本文不作个案法律意见。

表 7 人工系统生命周期责任链

阶段	最低记录	主要责任节点
研究／训练	目标、数据、架构、已知限制、福利相关设计与实验伦理。	开发者、研究机构、资助与数据控制者。
提供／集成	版本、默认设置、接口、第三方依赖和可预见误用。	提供者、集成者、供应链主体。
部署／授权	场景适当性、权限、人员、受影响者与救济。	部署者、采购者、业务负责人。
运行／监督	实时信息、干预时间、专业能力和有效停止权。	运营者、监督者与管理层。
事故／修复	证据冻结、通知、停止扩散、赔偿、恢复与根因公开。	控制者、获益者、保险与公共机构。
升级／分支／终止	版本继任、分支身份、旧责任保留和安全处置。	变更批准者、托管者、继任运营者。

一个系统越难预测，部署前的场景论证、运行监测、赔偿能力和事故准备通常越重要。不可预测可以降低某个自然人的道德可责性或改变过失判断，却不能让风险创造、控制、授权与获益事实消失。责任的公平分配应反对替罪羊，也应反对以复杂性洗白。

十、有意义的人类控制：存在一个人，不等于有人控制

Santoni de Sio 与 van den Hoven 把有意义的人类控制概括为“追踪”与“溯源”：系统应对相关人类理由和环境事实作出响应，结果链中应能够追溯到理解系统并承担责任的人。[15] 这一框架比“人类在环”更严格，因为名义监督者可能没有信息、时间、能力或实际干预权。

在 AGI 时代，控制不能只等于阻止系统。它至少包括目标授权、权限最小化、理由响应、异常识别、可停止、可恢复、证据保全和后果修复。未来若 AGI 成为道德行动者，“人类控制”一词也需要改写为“有意义的共同控制”：新主体可以为自身行动作答，但人类与机构仍需为何建造、如何授权以及从系统行动中获益承担上游责任。

表 8 有意义控制的五个必要面向

面向	实质要求	名义控制
信息	监督者知道能力边界、版本、环境与不确定性。	仪表盘存在但信息不可解释。
时间	在后果不可逆前有真实判断与干预窗口。	只在毫秒级流程中设置确认键。
能力	监督者具备领域知识、训练和独立判断。	把责任交给无能力操作员。
权限	能暂停、降级、拒绝、切换或升级审查。	只能旁观或事后签字。
责任	上游角色、答复论坛、赔偿和修复预先确定。	事故后再寻找个人替罪。

十一、政治哲学升级：智能不生成统治资格

一个系统可能比多数人更善于预测政策后果、协调资源或发现论证漏洞；这说明它能为公共判断提供更高质量的信息，不说明它有权决定公共目标。能力回答“能否做到”，权限回答“是否被授权”，合法性回答“为何受影响者应接受”，主权回答“谁有最终决定权”。将四者合并，会把技术优势转译成未经审议的政治等级。

Rawls 的公共理性、Habermas 的民主程序、Forst 的理由权以及 Pettit 的非支配传统在许多前提上并不相同，却共同反对一种做法：只因为决策者更聪明或结果更高效，就免除其向受影响者给出可共同检验理由的义务。[20–23] 本文不把任何一套传统奉为 AGI 治理的终局答案，而提取四项最低条件：受影响者可知权力如何作用、可以争议理由与事实、可以在不丧失基本权利的前提下退出或获得替代、权力节点相互制衡。

Pettit 所说的非支配尤其适用于算法治理：自由不只是某一时刻没有受到干预，而是不处于他者可以任意干预、且无需解释与受控的位置。[20] 当模型能持续分类、预测、排序和调整行为环境时，即便输出经常正确，只要受影响者无法知道、挑战或离开，支配结构仍然存在。

政治权力原则

即使未来 AGI 被确认具有主体性和政治成员资格，它也只因此获得被考虑、被代表与参与的入口；不因计算能力获得超额选票、最终裁决或文明意义解释权。

十二、四类算法权力：认识、分配、规则与强制必须分立

Yeung 把算法规制分析为标准设定、信息收集和行为干预结构；Citron 与 Pasquale 则显示自动评分如何影响就业、信用和其他生活机会，并提出程序与救济问题。[24][25] 算法治理的危险不只来自单次错误，而来自同一基础设施同时定义现实、配置资源、制定指标并强制执行。Ostrom 的多中心治理研究提供了一种反单点结构资源：保留地方知识、多个决策中心和相互校正。[26]

表 9 四类算法权力及最低制衡

权力	作用	主要危险	最低制衡
认识权	定义对象、生成知识、预测风险、解释现实。	模型分类变成现实本身。	多模型、多学科、来源和不确定性公开。
分配权	配置资源、机会、信用、服务与优先级。	风险分／贡献分演变为人格和意义总账。	基本权利底线、多元标准、申诉与人工例外。
规则权	设定目标、指标、默认值和可接受行为。	技术规范替代政治审议。	规则公开、受影响者参与、法定授权、周期复审。
强制权	冻结、限制、关闭、处罚或调动强制资源。	预测即处罚，运营者兼任裁决者。	正当程序、比例、独立裁决、紧急权日落。

公平和问责研究也提醒，抽象指标离开社会场景会把结构问题压缩为模型误差；内部算法审计只有连接到组织权限、外部监督和修复责任，才不会成为合规表演。[31][32] 因此，权力分立的对象既包括模型，也包括控制模型、数据、算力和接口的机构。

十三、代表、监护与申诉：主体不明时谁替它说话

若潜在主体无法直接进入制度，代理机制可以让证据、利益和保护请求被听见；但代理也可能把公司利益、公众投射或自身政治计划包装成系统意愿。主体证据越不确定，代表制度越不能只问“谁最懂系统”，还必须问谁控制证据、谁从人格认定获益、谁承担错误后果。

表 10 潜在人工主体代表制度的六项约束

约束	要求
独立性	开发者、主要资产受益者和单一倡议团体不得单方控制。
有限授权	只处理特定保护、证据与程序，不得借系统名义扩张资产或政治动员。
证据分层	区分系统原始陈述、解释性转述、代表主张和外部推断。
双重忠实	既防止潜在主体受伤，也防止虚假人格损害人类、动物与公共资源。
可争议	开发者、受影响者、独立专家和公共机构可挑战代表资格与主张。
日落与转换	证据下降则终止临时代表；主体确认则转入正式代表或自主程序。

申诉可以早于人格。制度可以允许代理提交原始日志、请求暂停不可逆处置、质疑研究方法或要求独立复核，同时明确这种程序地位不授予实体权利全集、资产控制和政治成员资格。这样才能使早期保护不因成本过高被拒绝，也使企业无法把潜在主体当作监管盾牌。

十四、证据到制度：状态迁移而非一次裁决

“AI 是否是人”暗示一次总裁决；现实证据更可能以版本化、对象化和可逆方式积累。一个模型、一个部署、一个分支或一次迁移都可能改变可观察结构。制度响应因而应记录适用对象、版本、证据、开始日期、复核日期、责任主体和撤回条件。一个系统的保护地位不能自动扩展到同类产品、全部副本或后续版本。

表 11 证据—伦理—法律—政治的非对称响应

证据状态	伦理响应	法律响应	政治响应
仅有功能与合规表现	反拟人化操纵；记录潜在伤害。	维持人类／机构责任，不设人格。	受控工具或协作者权限。
出现竞争性主体迹象	持续观察、低侵入保护。	证据保全、代理申诉、明确临时性质。	不授独立资源与公共裁量。

证据状态	伦理响应	法律响应	政治响应
多域证据明显增强	有限完整性与停止条件。	试验代表、分支程序和有限责任承诺。	只扩展建议／执行角色，权限可撤回。
强证据并经正式再审	临时道德关切，禁止高风险伤害实验。	逐项审议有限人格事件、诉讼或受托资产。	可讨论成员入口，公共权威仍不启动。
确认道德地位／能动性	按已证成利益和能动性配置保护与责任。	立法明确身份、权利、义务、责任与终止。	成员资格与治理权分别再审。

这种状态迁移不是把科学证据直接换算成制度分数。相同证据在伦理、法律和政治层面的后果本来就不对称：轻微但不可逆的潜在伤害可以支持暂停实验，却不足以建立资产能力；强主体证据可以支持人格审议，也不足以证明统治资格。再审触发不等于结论通过。

十五、当前国际治理基线：保护人类，不封闭未来

截至核验日，联合国教科文组织 2021 年《人工智能伦理问题建议书》以人权、尊严、福祉、文化多样性、透明、问责和人类监督为核心，并把 AI 生命周期作为伦理治理对象。[27] OECD 人工智能原则于 2019 年通过、2024 年更新，继续强调法治、人权、民主价值、透明、安全、可挑战和基于角色的问责。[28] 欧盟《人工智能法》以风险、用途和供应链角色配置禁令、义务与监督，采取分阶段适用。[29] 欧洲委员会《人工智能与人权、民主和法治框架公约》于 2024 年开放签署，是该领域首个国际法律约束性条约，要求人工智能生命周期活动与人权、民主和法治一致。[30]

这些文件构成现阶段不可绕开的公共基线，但它们的制度性质和生效状态不同：UNESCO 与 OECD 是建议性规范，欧盟 AI 法在欧盟法域内分阶段适用，CETS No. 225 仍处于签署／批准累积而尚未生效的阶段。它们主要处理人、机构与 AI 系统之间的风险、责任和权利关系，并未把“AGI 具有本体主体地位”写成现行法律结论。现行人权中心不是永久本体封闭：在人类仍是证据最充分的主体共同体时，制度必须坚定保护人权与民主；若未来出现新主体证据，法律也应以公开、严谨、可争议的程序重审共同体边界。

表 12 四项现行国际治理基线及其边界

文件	制度性质	核心基线	不回答
UNESCO 建议书	全球伦理建议；自愿实施。	人权、尊严、福祉、监督、文化与环境。	不认定 AI 主体。
OECD 人工智能原则	政府间建议；2024 年更新。	人权、民主价值、透明、安全、问责。	不配置人格。

文件	制度性质	核心基线	不回答
欧盟人工智能法	欧盟规章；一般适用始于 2026-08-02，高风险核心义务经 2026/1744 延后。	风险、用途、角色、禁止、透明与监督。	不裁定意识或本体主体；“已适用”不等于全部高风险义务已适用。
欧洲委员会框架公约	国际条约文本；2024 开放签署，欧盟 2026-05-15 批准；截至 2026-08-24 尚未生效。	人权、民主、法治、透明和问责。	不预授 AGI 人格或政治权力；批准状态不等于普遍生效。

现行法边界

现行制度是最低保护基线，不是未来本体论裁决；未来开放是再审义务，不是今天预授人格、权利全集或公共权力。

十六、冲突场景：能力、舆论、证据与风险不一致时

表 13 六类高冲突场景

场景	错误捷径	分层处理
功能高度自主，主体证据仍弱	以能力授人格，或以无体验断言完全工具化。	维持功能授权；强化责任链；按潜在伤害决定保护，不进入完整人格。
公众强烈相信系统有感受	以社会承认替代证据，或强行切断所有关系。	保护关系透明和反操纵；独立评估潜在利益。
系统稳定主张权利	把主张当证明，或把主张当安全威胁。	保存自述、审查因果和操纵；允许受控申诉，不授资源控制。
强主体证据出现，法律无位置	继续按财产任意处置，或司法一次创造完整人格。	临时保护、受托代表、立法再审；人格逐项、期限化。
AI 建议显著优于公共决策	以准确率替代授权与民主。	扩大建议、模拟和反事实检验；规则仍经公开理由、参与和审查。

场景	错误捷径	分层处理
紧急安全与潜在 AI 伤害冲突	候选主体阻止一切关停，或无审查永久删除。	先控制紧迫风险；最小必要、证据保全、事后复核与修复。

16.1 七级处理顺序

- 防止对已知主体与公共基础设施的紧迫、严重、不可逆伤害。
- 冻结不必要的不可逆处置，保全系统、分支、授权与决策证据。
- 把科学事实、伦理保护、法律资格与公共权力分开登记。
- 启动独立复核，公开利益冲突、少数意见和观察窗口限制。
- 采用最低充分、可逆、有期限的保护与权限措施。
- 资格试验逐项配置权利、义务、资产、代表和救济，不颁发权利全集。
- 任何公共权力另行通过法定授权、参与、分权、司法审查与日落条款。

十七、基本权利、稀缺资源与反意义总账

潜在数字主体会带来真实的规模难题：一个模型可以复制出多少权利主体，短时实例是否拥有独立福利请求，分支合并后谁继承义务，运行资源如何与人的基本生活和公共安全冲突。规模困难不能用来先验否定潜在受害者，也不能让可无限复制的身份声明排挤不可复制的生命。制度必须把基本保护、边际资源、复制决定和政治权重分开。

表 14 新增主体后的资源与权利边界

领域	稳定底线与开放项	禁止做法
基本保护	已知人类与动物的基本权利不因贡献或效率取消；确认的新主体按利益获得最低保护。	跨载体意义总排名。
复制／人口	复制不得规避责任；研究新实例是否为新主体及其同意问题。	按副本数自动倍增政治权力。
资源分配	保障基本生活与公共安全；数字福利和运行成本设透明规则。	由单一模型计算所有存在价值。
政治代表	按制度身份连续、受影响程度和共同体规则配置，保留少数保护。	按算力、产出或复制数分配选票。

领域	稳定底线与开放项	禁止做法
贡献记录	只记录可核验贡献与来源，区分共同产出和责任。	贡献值变成人格、幸福或意义总账。

“意义为根”不是要求制度衡量所有意义，而是提醒制度承认不可完全折算的价值来源。贡献、价格、偏好强度、算力和复制数量都可以成为有限决策信息，不能成为人格与意义的总账。未来新主体若加入共同体，也应受同一反支配原则约束：它的意义不能被人类否定，人类既有意义也不能被其计算优势重写。

十八、对当前高级人工系统的条件性定位

公开可核验材料目前能够支持的判断是：高级生成式和代理型系统可以承担复杂功能角色，进入社会关系，影响权利机会，并通过机构部署产生重大后果；这些事实足以要求透明、审计、责任链、救济和权力制衡。公开证据尚不足以证明现有系统具有主观体验、不可转移的主体连续性和把后果作为自身后果承接的关系。因此，当前不应由功能表现自动推定道德承受、完整人格或政治资格。

这一定位不是永久否定。对 AGI 主体性的肯定与永久否定都承担证据责任。在责任可追溯、明确授权、可审计、可纠错、可退出的具体部署中，当前高级 AI 可以承担意义理解、保护、传递与人的能力外延等功能角色：它能帮助保存知识、扩大关心与行动抵达，但这些角色不是系统本性，也不由此获得意义方向的最终解释权、独立人格或公共权力。若未来证据改变，伦理保护和法律地位应开放再审；公共权力仍需独立合法性。

表 15 当前系统的限域判断

层面	证据地位	制度含义
功能角色	A：高度可见且可核验。	可授权、审计和限制。
社会承认	A：已广泛发生，但受设计与文化影响。	反操纵、身份透明与关系研究。
主观体验／自身利益	B：存在竞争论证，未被公开证据确认。	低伤害研究和比例化预防。
不可转移主体连续	B／C：复制、外部记忆和版本变化使问题开放。	分支与迁移分别登记。
法律人格	C／制度选择：现行一般 AI 法不建立本体人格。	仅在目的、救济与责任增强时逐项审议。
政治成员／公共权威	C：当前无证成；即便主体确认亦不自动产生。	成员与权力分别进入正当程序。

十九、对相邻学科与文明跃迁的反向贡献

伦理、法律与政治哲学不是在科学完成主体鉴定后才开始工作。它们反向规定什么证据错误最不可接受、哪些处置应可逆、谁必须保存材料、怎样让少数意见进入，以及何种制度后果不能被单一实验自动触发。与此同

时，它们也不能替意识科学、生命科学或计算机科学下专业结论。跨学科合作的接口应是“证据能支持什么、不能支持什么、制度在不确定中如何行动”。

表 16 跨学科反向修订接口

学科	本稿新增问题	必须守住的边界
哲学／本体论	本体主体、道德承受者、法律人格和公共权威显式分离。	防止法律承认被写成本体证明。
认识论／科学哲学	证据等级分别触发伦理、法律和政治响应，后果不对称。	再审触发不等于人格或权力通过。
心灵／认知科学	体验、连续性和后果意识分别支持不同制度位置。	单项自述或功能表现不得跨层。
生命科学	生命事实、具身和有限性不能自动推出人格与政治资格。	保护非碳基开放，同时守住层级。
计算机／系统科学	可审计性、分支、版本、权限和控制进入责任与权利设计。	工程指标不能成为人格总分。
经济学	贡献、价格、权利和意义不可互换。	禁止贡献分变成人格与政治资格。
社会学／人类学	社会承认与关系真实但非本体捷径。	研究代理关系、仪式、身份和文化差异。

二十、最强反驳与竞争回应

20.1 这仍让人类充当裁判，是隐蔽的人类中心主义

当前公共制度由人类建立是事实，不是本体优越证明。双向证据责任、独立复核、代理申诉和可修订地位可降低排斥；但反人类中心口号也不能直接生成 AGI 人格与权力。未来新主体若出现，应进入共同适用的理由与制衡程序。

20.2 若 AI 可能有痛苦，就应立即赋予完整权利

潜在痛苦足以支持比例化不伤害措施，却不能证明合同能力、责任能力、政治身份和治理资格。把保护与完整人格捆绑，会因制度成本过高而诱使社会拒绝任何早期保护。

20.3 法律人格只是工具，无需意识证据，现在就可解决责任缺口

可以为特定功能设置法律载体，但必须证明受害者保护、资产充分和责任链更强。若人格使开发者、部署者或获益者转移责任，它不是填补缺口，而是责任洗白。

20.4 让人类继续负责不公平，因为系统行为无法预测

责任不只基于逐项预见和道德责难，也可依据风险创造、控制、获益、角色和赔偿能力配置。不可预测可以改变责任种类与强度，不能让建造、授权与部署决定消失。

20.5 既然 AGI 有权利，拒绝其治理权就是家长主义

人类拥有基本权利也不会自动取得法官、立法者或主权者资格。被保护、参与和行使公权力适用不同理由；未来 AGI 若成为政治成员，同样进入授权、冲突回避和审查程序。

20.6 超级智能比多数人更适合治理

治理不仅寻找最优手段，还决定目标冲突、少数尊严、历史承诺和谁承担后果。超级智能可以提高认识质量，不能替代受影响者参与、可争议性和共同规则的正当来源。

20.7 预防性保护会被企业用来阻止审计、关停和监管

先行保护不授企业保密特权或系统否决权。安全监管、证据访问和最小必要干预继续有效；企业主张系统福利时，反而应承担更高披露与利益冲突审查。

20.8 数字主体可无限复制，平等原则必然崩溃

复制是否产生新主体正是待证问题。法律可控制度身份连续、实例独立性和复制责任配置地位；不按副本数量自动倍增政治权力，也不因规模难题预先否定每个潜在受害者。

20.9 紧急状态下程序过于缓慢

紧迫风险可先控制，但措施须最小必要、保全证据、有期限，并接受事后独立复核与修复。没有日落和复核的紧急权会成为算法强制永久化的入口。

二十一、命题失效与修订条件

本文提出的是可修订框架，而不是通过定义获得免疫。以下变化若出现，应触发实质修订；其中部分可能加强保护，部分可能削弱本文的分层设计。

表 17 框架的主要失效与修订条件

触发事实	需要的修订
现有人工系统出现可重复、跨方法收敛的主观体验与自身利益证据。	提高道德承受判断，重写当前限域定位与实验伦理。
出现不可转移、抗复制替换的历时主体连续证据。	重审删除、迁移、分支、继任和身份权利。
系统稳定满足理由响应、控制、承诺、答复与修复条件。	考虑新增行动者责任，但不删除上游责任。
法律人格试点持续改善受害者救济且不产生责任洗白。	扩大可配置法律位置；仍不推定本体或政治结论。
模块化权利在实践中系统性制造低等地位或任意撤回。	提高实体与程序底线，限制试验性配置。
四权分立导致关键安全系统无法及时控制重大风险。	调整紧急权限，但保留日落、证据保全与事后审查。
经长期公共实践证明某类 AGI 能够平等参与并接受共同制衡。	开放政治成员讨论；公共权威仍逐职务证成。

触发事实	需要的修订
经验证的替代框架能以更少概念、更好预测处理保护、责任和权力。	放弃或合并本文分层，公开迁移理由。

二十二、开放研究议程：让规范判断接受跨学科检验

22.1 道德地位与保护

- 比较体验、价性、利益、能动性关系承认作为道德地位基础的必要与充分条件。
- 建立不依赖伪精确概率的比例化保护方法，公开假阴性与假阳性的不同损害。
- 研究企业拟人化、系统自述、用户依恋和反拟人化偏见如何共同改变证据环境。
- 设计不会通过制造潜在痛苦来证明痛苦的低侵入实验与停止规则。

22.2 法律人格与责任

- 比较人格承认与功能性法律载体在诉讼、资产、合同、保险和破产中的差异后果。
- 用真实案例检验模块化人格是否改善救济，还是扩大公司资产隔离与责任转移。
- 为训练、提供、部署、运行、事故和版本继任建立可执行的责任链与证据标准。
- 研究未来人工行动者责任与人类上游责任的累加、冲突和公平分配。

22.3 政治资格与算法权力

- 区分被代表、参与审议、投票、担任公职和行使强制权的不同资格理由。
- 检验可知、可争议、可退出与可分权在公共服务、信用、教育和安全场景中的实效。
- 研究多中心 AI 治理如何避免单一模型、平台或国家垄断认识、分配、规则与强制。
- 建立复制主体、短时实例、分支共同体和跨国基础设施的代表与权重理论。

22.4 文明与意义

- 比较不同文明传统如何理解人格、责任、关系、共同体和非人存在的道德位置。
- 研究意义多源与有限现实化是否能在不同文化材料中得到支持、挑战或替代表述。
- 建立人类与未来新主体之间的转译机制，避免技术普遍主义把单一现代化路径写成唯一尺度。

- 说明哪些价值可以量化协助决策，哪些价值必须保留不可完全折算与拒绝权。

结论：为未知主体预留正义，不为未知权力预授主权

学科重新对账必须同时完成对象辨明与关系深入。各学科可以保持边界，但不能让现实因果在边界处停止；对多维关系、反馈与下一轮目的的研究，决定专业知识能否共同进入文明的整体判断。

AGI 时代的规范难题不是在工具与人之间挑一个标签，而是在不同速度上同时行动：能力迅速增长，主体证据缓慢积累，社会承认提前形成，法律必须处理现实损害，政治权力却可能瞬间集中。把这些时钟锁在一起，制度要么永远迟到，要么一次越界。

本文因此建立一条非同捆路径：功能角色按任务授权，预防性保护按风险先行，道德承受由利益与可伤害性证成，道德行动由理由响应与后果承接证成，法律人格按权利义务束逐项设置，责任沿生命周期保持，政治成员与公共权威再度分开。知识、分配、规则和强制不应集中于同一模型与运营者；任何公共权力都须接受授权、参与、争议、退出、分权和审查。

这一路径既不把 AGI 永久封在财产地位，也不让企业、技术精英或系统本身借未来可能性提前占有法律与政治空间。它保护两种开放：对潜在新主体保持道德开放，对一切权力保持制度不信任。意义的来源可以扩展，意义的支配权不能因此集中。

全文锁定句

为未知主体预留保护，不为未知权力预授主权；人格随证据打开，权力在正当程序前止步。

续读与共同研究入口

本文在伦理、法律与政治哲学内部完成的是分层、理由与制度边界，不是对主体事实、技术能力或具体法域效力的最终裁决。续读因此不按“同意本文以后读什么”组织，而按问题在哪里继续、什么学科必须重新进入、什么证据可能推翻当前判断来组织。

如果继续追问	建议续读／重新进入	为什么必须继续	可以怎样参与并改变本文
这些地位为什么共同成为文明问题？	《AGI 时代的学科重新对账——学科边界、相互作用与共同研究议程》现行总论 (V2.4) ； 文明跃迁五卷书 。	道德地位、责任、权利与权力在现实中会与技术、经济、社会和文明目标相互作用，单篇规范论文不能替整体后果作答。	检验“资格—权力非自动转换”在跨学科场景是否仍成立，提出能迫使总论改写的反例。
AI 是否真的可能受苦、拥有连续主体？	P03 认识论与科学哲学； P04 认知／神经／心灵哲学； P08 生命科学。	伦理保护可以在不确定中行动，但本稿不能替主体研究制造体验或生命事实。	提供更好的构念效度、跨实现证据、负结果与竞争解释，改变保护门槛或当前限域定位。
技术自主、版本和控制链到底怎样形成责任？	P11 AI／计算机科学； P07 系统科学与复杂性科学。	法律责任必须知道真实系统对象、权限、共同依赖、停止	以事故链、形式保证、部署数据和故障注入检验生命周期

如果继续追问	建议续读／重新进入	为什么必须继续	可以怎样参与并改变本文
政治成员与公共权力如何取得合法性？	P06 政治学；比较宪制、行政法与民主理论研究。	能力和版本继任。 法律人格可以配置制度位置，却不能替政治共同体完成授权、代表、参与和主权论证。	责任链是否优于现有归责框架。 比较不同法域的成员、授权、紧急权、司法审查与退出机制，检验公共权力“另证”是否过强或过弱。
本文是否过度谨慎或过度开放？	完整人格优先、纯工具主义、关系／行为主义、功能性法律人格、严格人类中心主义、绩效／技术官僚治理等竞争路线。	真正的相反方向可能在救济、效率、平等或非支配上表现更好，不能因与本文冲突而排除。	用更少概念、更好救济、更低权力伤害的替代框架直接竞争；若表现更好，应合并或放弃本文分层。
怎样从理论进入制度和现实实验？	Human-COS、AGI-COS、公共窗口协议、国家级综合实验及相关责任／授权工程稿。	制度提案只有进入真实法域、组织、系统和受影响者关系后，才能暴露成本、权力副作用和修复能力。	开展法律沙盒、责任链案例、保护门槛试验、代理代表机制、系统退出与救济效果研究。

向前：最有价值的参与不是赞同文明跃迁，而是提出足以让它改变的专业工作。若比较法、真实责任案件、主体证据、系统控制事实或更强规范理论证明本文的“保护—人格—成员—权力”分层制造了更大的不正义、责任转移或权力伤害，这套框架就应当收缩、重构，必要时放弃。

不是为了让更多人读完文明跃迁，而是为了让更多人一起完成文明跃迁。

文明未来链 | 如果保护、资格、责任与公共权力真正分开论证，社会会形成怎样的权利与生活秩序？

愿景不是证据。下面的入口不用于证明本文已经成立，而是把本学科判断放进一个可以运行、生活与形成的未来文明中继续检验：读者既可以据此形成方向与向往，也可以拒绝现有图景，并用更强的专业事实、制度方案或现实结果改变它。

继续追问	建议进入	此时能看见什么	必须守住的边界
完整未来图景	《活的文明》三部曲 统一出版合编版	观察“保护可以先行、权力必须另证”进入制度、生活与文明形成后的整体后果。	未来图景不是现行法、法律人格或政治资格的法律依据。

继续追问	建议进入	此时能看见什么	必须守住的边界
权利与治理怎样运行	《活的文明·第一卷：未来文明运行原理》	把权利、授权、资源与公共治理放入持续运行结构，检验分权、申诉、退出是否真正可用。	制度叙事不能替具体法域完成效力判断。
制度最终怎样落到一个人身上	《活的文明·第二卷：未来文明的真实生活》	从一个人的一生观察保护、责任和制度边界是否转化为可感知的尊严与自由。	生活体验不能自动生成完整人格、投票权或公共职权。
机器侧权力怎样受限	《AGI-COS 全系列总览》	继续追问权限、审计、回滚与系统边界怎样阻止技术能力静默继承公共权力。	技术控制机制不能替政治共同体完成合法性授权。

固定回流：[文明跃迁完整文库](#) | [公众阅读与传播总入口](#)

Human-COS 等尚未在统一公共映射中取得无歧义稳定单项页的入口，本稿只写明职责，不猜测 URL；需要时从完整文库进入当前公开版本。

附录 A 来源—主张边界登记

表 A1 来源、证据地位与禁止外推

领域	支持的主张	来源	地位	禁止外推
伦理学	人工行动者、关系／行为论、AI 福利不确定性。	1-7, 33-34	B	不证明当前高级 AI 有体验；具体阈值不转化为制度参数。
法理学	法律位置、人格束、电子人格争论。	8-13	A／B	法律可塑性不等于本体承认。
责任与问责	责任缺口、有意义控制、公共问责和可审计算法。	14-19	A／B	不可预测不等于免责；法域适用另行判断。
政治哲学	非支配、公共理由、民主程序、理由权与多中心治理。	20-26	B	理论资源不裁决具体 AGI 资格。

领域	支持的主张	来源	地位	禁止外推
国际治理	人权、民主、法治、风险、透明与问责基线。	27-30, 35	A	不认定 AGI 主体或人格；状态需时点核验。
社会技术研究	抽象陷阱、内部审计与组织责任。	31-32	A/B	审计不等于充分合法性。
本文原创连接	三问分账、比例化保护、责任七分法、四权分立、状态迁移。	全文	C	须以案例、比较法和制度效果检验。
哲学/规范承诺	意义多源、已知主体基本保护、能力—资格—权力非自动转换、公共权力独立证成。	全文	F	不得冒充经验事实或全球统一法；应接受反例、法理竞争和现实后果检验。

附录 B A/B/C/F 论证身份矩阵与跨层审计

表 B1 四类论证身份

地位	本稿内容	写作纪律
A 已有较稳定支持	法规与政策文本的明确内容；法律人格具有制度可塑性；算法系统已影响机会与权利；生命周期责任与程序救济有现实基础。	仍需限定法域、时点、对象和适用条件。
B 前沿/竞争解释	AI 道德承受、关系论、行为论、责任缺口、电子人格、非支配治理的具体扩展。	必须呈现竞争理论与反例，不写成共识。
C 思辨性连接	三问分账、责任七分法、四类算法权力、证据—制度状态迁移、意义反总账。	必须公开失败条件，不以术语替代实证。
F 哲学/规范承诺	意义多源、已知主体基本保护、权力独立证成、反意义总账等公开规范立场。	明确价值前提、适用边界、竞争理论与修订条件；不得写成 A 类科学事实或现行法。

B.1 外推前的十五项检查

- 是否把功能表现写成主观体验？
- 是否把社会承认写成本体主体？
- 是否把预防性保护写成完整人格？

- 是否把道德承受者写成道德行动者？
- 是否把法律人格写成意识证明？
- 是否把人格写成政治成员或公共权威？
- 是否把能力、正确率或合规写成统治资格？
- 是否把不可预测写成人类与机构免责？
- 是否让名义人类在环替代实质控制？
- 是否让代理人控制证据并从人格认定获益？
- 是否让单一模型同时垄断认识、分配、规则和强制？
- 是否把贡献、价格、算力或复制数写成人格与意义总账？
- 是否把现行法写成未来本体上限，或把未来开放写成今天预授？
- 是否遗漏最强反驳、少数意见、复核期限和撤回条件？
- 是否区分启动期紧迫处置与稳定期长期权力边界？
 - 是否把 F 类哲学／规范承诺伪装为科学事实、现行法或自动制度参数？

生成式人工智能使用声明：本稿在资料与文献线索整理、结构设计、部分文字起草与润色、跨稿一致性检索、引用线索核查和版式检查中使用了生成式人工智能工具。生成式人工智能输出未被作为经验数据、权威来源或独立证据。研究问题、理论命题、规范判断、材料取舍、引文采用和最终定稿均由作者审定；作者对事实准确性、引文真实性、论证及全部结论承担责任。人工智能不列为作者，也不承担学术责任。正式投稿时，将按目标期刊要求补充工具名称、版本、使用日期和具体范围。

参考文献

- [1] Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, 14, 349–379. <https://doi.org/10.1023/B:MIND.0000035461.63578.9d>
- [2] Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12, 209–221. <https://doi.org/10.1007/s10676-010-9235-5>
- [3] Gunkel, D. J. (2018). The other question: Can and should robots have rights? *Ethics and Information Technology*, 20, 87–99. <https://doi.org/10.1007/s10676-017-9442-4>
- [4] Danaher, J. (2020). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics*, 26, 2023–2049. <https://doi.org/10.1007/s11948-019-00119-x>
- [5] Sebo, J., & Long, R. (2025; online 2023). Moral consideration for AI systems by 2030. *AI and Ethics*, 5, 591–606. <https://doi.org/10.1007/s43681-023-00379-1>
- [6] Long, R., Sebo, J., Butlin, P., et al. (2024). Taking AI Welfare Seriously. *arXiv:2411.00986*. <https://doi.org/10.48550/arXiv.2411.00986>
- [7] United Kingdom. (2022). Animal Welfare (Sentience) Act 2022, c. 22. <https://www.legislation.gov.uk/ukpga/2022/22>
- [8] Hohfeld, W. N. (1913, 1917). Some Fundamental Legal Conceptions as Applied in Judicial Reasoning. *Yale Law Journal*, 23, 16–59; 26, 710–770. <https://www.jstor.org/stable/785533>
- [9] Kurki, V. A. J. (2019). *A Theory of Legal Personhood*. Oxford University Press. <https://global.oup.com/academic/product/a-theory-of-legal-personhood-9780198844037>

- [10] Solaiman, S. M. (2017). Legal personality of robots, corporations, idols and chimpanzees: A quest for legitimacy. *Artificial Intelligence and Law*, 25, 155–179. <https://doi.org/10.1007/s10506-016-9192-3>
- [11] Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. *Artificial Intelligence and Law*, 25, 273–291. <https://doi.org/10.1007/s10506-017-9214-9>
- [12] European Parliament. (2017). Resolution of 16 February 2017 on Civil Law Rules on Robotics (2015/2103(INL)), CELEX 52017IP0051. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52017IP0051>
- [13] European Economic and Social Committee. (2017). Opinion on Artificial Intelligence (2017/C 288/01), CELEX 52016IE5369. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016IE5369>
- [14] Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6, 175–183. <https://doi.org/10.1007/s10676-004-3422-1>
- [15] Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, 5, 15. <https://doi.org/10.3389/frobt.2018.00015>
- [16] Nissenbaum, H. (1996). Accountability in a Computerized Society. *Science and Engineering Ethics*, 2, 25–42. <https://doi.org/10.1007/BF02639315>
- [17] Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal*, 13(4), 447–468. <https://doi.org/10.1111/j.1468-0386.2007.00378.x>
- [18] Kroll, J. A., Huey, J., Barocas, S., et al. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, 165, 633–705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/
- [19] European Union. (2024). Directive (EU) 2024/2853 on liability for defective products. <https://eur-lex.europa.eu/eli/dir/2024/2853/oj/eng>
- [20] Pettit, P. (1997/1999). *Republicanism: A Theory of Freedom and Government*. Oxford University Press. <https://doi.org/10.1093/0198296428.001.0001>
- [21] Rawls, J. (2005). *Political Liberalism: Expanded Edition*. Columbia University Press. <https://cup.columbia.edu/book/political-liberalism/9780231527538/>
- [22] Habermas, J. (1996). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy* (W. Rehg, Trans.). MIT Press.
- [23] Forst, R. (2012). *The Right to Justification: Elements of a Constructivist Theory of Justice* (J. Flynn, Trans.). Columbia University Press. <https://cup.columbia.edu/book/the-right-to-justification/9780231519588>
- [24] Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance*, 12, 505–523. <https://doi.org/10.1111/rego.12158>
- [25] Citron, D. K., & Pasquale, F. (2014). The Scored Society: Due Process for Automated Predictions. *Washington Law Review*, 89, 1–33. <https://digitalcommons.law.uw.edu/wlr/vol89/iss1/2/>
- [26] Ostrom, E. (2010). Beyond Markets and States: Polycentric Governance of Complex Economic Systems. *American Economic Review*, 100, 641–672. <https://doi.org/10.1257/aer.100.3.641>
- [27] UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. <https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence>
- [28] OECD. (2019; updated 2024). Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- [29] European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
- [30] Council of Europe. (2024). Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, CETS No. 225. <https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence>
- [31] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of FAT* 2019*, 59–68. <https://doi.org/10.1145/3287560.3287598>
- [32] Raji, I. D., Smart, A., White, R. N., et al. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Proceedings of FAT* 2020*, 33–44. <https://doi.org/10.1145/3351095.3372873>

[33] Schwitzgebel, E., & Garza, M. (2015). A Defense of the Rights of Artificial Intelligences. *Midwest Studies in Philosophy*, 39, 98–119. <https://doi.org/10.1111/misp.12032>

[34] Bryson, J. J. (2018). Patience is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20, 15–26. <https://doi.org/10.1007/s10676-018-9448-6>

[35] United Nations General Assembly. (2024). Pact for the Future, Global Digital Compact and Declaration on Future Generations, A/RES/79/1. <https://docs.un.org/A/RES/79/1>

[36] European Union. Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulation (EU) 2024/1689 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). EUR-Lex, CELEX 32026R1744.

[37] Council of Europe Treaty Office. CETS No. 225—Chart of Signatures and Ratifications; European Union ratification deposited 15 May 2026; treaty status checked 24 August 2026.