对外经济贸易大学法学院

博士申请学位科研成果考核折算科研文献

文献形式：文献综述

文献题目：高风险人工智能的法律治理研究

博士论文题目：高风险人工智能的法律治理研究

导师：梅夏英

作者姓名：王灵娟

专业：民商法学

学号：202300520109

高风险人工智能的法律治理研究文献综述

摘要：高风险人工智能法律治理法案是我国人工智能立法领域的重要议题。由于我国目前尚无明确的高风险人工智能的法律监管规则，该议题有探讨必要性。根据 AI 造成不同的风险类型提供不同的法律调整方案，对于风险级别和类型较低 AI 施加更为宽松的法律，高风险 AI 应施加更严格的法律措施。此为目前全球比较公认的人工智能治理路径。学界对高风险 AI 的治理研究分为高风险 AI 的技术类型和风险表现形式、高风险 AI 的正面效应，高风险 AI 的理论基础、高风险 AI 的运用范畴、高风险 AI 运用的边界以及既有高风险 AI 具体治理路径研究六个主要部分。高风险 AI 的技术类型复杂多样，但 AI 技术的复杂度与风险并不呈现正相关关系，技术简单和技术复杂的 AI 都可能造成高风险。高风险 AI 具有正面效益，所以我国的高风险 AI 的法律治理应坚持包容审慎的原则，在促进产业发展和控制风险之间找到平衡点。高风险 AI 的理论基础主要来源于哲学和社会学中的伦理基础。高风险 AI 的监管规则主要分为一般监管规则和特殊监管规则，两类规则都具有局限性。高风险 AI 的私法规则还有拓展和改良的空间。前述的既有学术研究表明，高风险 AI 法律治理的构建方案可通过完善判定机制，改善监管规则，发展私法路径展开。

关键词：人工智能；高风险；私法规则；监管规则；风险路径；

高风险人工智能法律治理是我国人工智能立法领域的重要议题。截至目前，我国虽尚未在基本法律位阶上出台《人工智能法》，根据现有学术研究以及较为分散或者层级相对较低的各种规范性法律文件可以推论出人工智能包容审慎的基本立法动向。与此同时，风险规制路径是不能被忽视的立法路径。问题在于，如何让高风险 AI 法律治理在促进产业发展和控制风险之间寻求平衡。高风险 AI 概念该如何界定和识别，如何能设计出较为稳定又能适应高风险 AI 技术发展的法律框架，是值得研究的学术问题。厘清既有高风险 AI 现有研究现状，可为高风险 AI 法律治理研究提供研究基础与创新指引。关于高风险人工智能法律治理的文献综述拟从六个方面展开：高风险 AI 的技术类型与风险表现形式、高风险 AI 的正面效益，高风险 AI 的法律判定机制、高风 AI 的治理理论、高风险 AI 的监管规则，高风险 AI 的私法规则。

一、高风险 AI 的技术类型与风险表现形式

AI 是人类无法全面掌控其输出结果，其也无法全面实现人类意图的智能程序。AI 和人类意图始终存在匹配度的差异性。高风险 AI 意味着其输出结果可能让人类无法接受或者无法预测。从纯粹理学角度分析，AI 技术可以抽象为物理学、数学、计算机学科中的基本理论，如概率、线性代数、微积分、编程、代码等。[①]AI 技术以“深度学习”或者“机器学习”为核心技术机理。深度学习和机器学习有促进 AI 变化和进步的可能，但还是会局限于形式化的计算、逻辑与理性方面的表征。”[②] 自然科学发展决定了 AI 技术发展上限。AI 技术的分类分级和发展路线都较为复杂，在纷繁复杂的 AI 实践技术背景之下，判定哪类技术会构成高风险 AI，实属困难，原因亦在于：AI 技术是否复杂与是否造成高风险没有本质联系。

(一)高风险 AI 的技术类型

1.离身智能、具身智能、反身智能、共生智能

在 AI 产业理论界，AI 有四种主要的发展形态。离身智能（Disembodied Intelligence）是指通过符号运算、逻辑推理或数据驱动模型实现的智能形态，它假设智能可以独立于身体存在，依赖大脑或算法的内部处理机制，无需通过身体与环境的实时交互来塑造认知过程。早期 AI 研究者认为，机器可以通过对符号的操作实现对人类思考(理性行为)的模拟，即让机器像人类那样思考，无需考虑身体和环境。他们把人工智能定义为模拟人类的智能行为，除了模拟思考之外，

[①] 杜文静、刘海：《贝叶斯人工智能的概率证成》，载《科学技术哲学研究》2022 年第 4 期，第 19 页。
[②] 董春雨：《从机器认识的不透明性看人工智能的本质及其限度》，载《中国社会科学》2023 年第 5 期，第 166 页。

还包括感知、动作，以及情感与灵感等。离身智能不会没有考虑模拟人类行为之间的相互联系。

具身智能(Embodied Artificial Intelligence，EAI)是指一种基于物理身体进行感知和行动的智能系统，其通过智能体与环境的交互获取信息、理解问题、做出决策并实现行动，从而产生智能行为和适应性。具身智能包括四大核心要素，即物理本体、智能体、数据和学习进化框架。具身智能是 AI 的重要分支，是智能机器人走向更高阶应用的主要途径，人形机器人则是智能机器人的最高形态。[①]

反身智能（Reflexive Intelligence）是指智能系统通过自我监控、反馈调节和元认知能力实现动态优化的智能形态。这一概念源于控制论中的负反馈机制以及认知科学中的元认知理论（Metacognition），强调智能不仅是对外界的反应，更是对内部状态的持续修正。

共生智能（symbodied Al）共身智能是指通过技术手段将人类智能与以人类视角为基准的 AI 进行深度对齐和融合，同时整合各自物理载体 (人体 + AI 硬件) 与环境动态交互，形成以人为中心的协同感知、高效决策与交互执行能力。这种智能形式不仅强调人类与 AI 之间在信息层面的双视角协同， 还包括在物理层面的整合。

2.嵌入模式 （Embedding）、副驾驶模式（Copilot）和智能体模式（Agent）

从人机协同角度又可以将 AI 分为嵌入模式 （Embedding）、副驾驶模式（Copilot）和智能体模式（Agent）。嵌入（Embedding）模式，用户通过与 AI 进行语言交流，使用提示词来设定目标，然后 AI 协助用户完成这些目标，比如普通用户向生成式 AI 输入提示词创作小说、音乐作品、3D 内容等。在这种模式下，AI 的作用相当于执行命令的工具，而人类担任决策者和指挥者的角色。

副驾驶（Copilot）模式，在这种模式下，人类和 AI 更像是合作伙伴，共同参与工作流程，各自发挥作用。AI 介入到工作流程中，从提供建议到协助完成流程的各个阶段。例如，在软件开发中，AI 可以为程序员编写代码、检测错误或优化性能提供帮助。人类和 AI 在这个过程中共同工作，互补彼此的能力。AI 更像是一个知识丰富的合作伙伴，而非单纯的工具。

智能体（Agent）模式，人类设定目标和提供必要的资源（如计算能力），然后 AI 独立地承担大部分工作，最后人类监督进程以及评估最终结果。这种模式下，AI 充分体现了智能体的互动性、自主性和适应性特征，接近于独立的行动

[①] Pfeifer, Rolf and Fumiya Iida. “Embodied Artificial Intelligence: Trends and Challenges.” Embodied Artificial Intelligence (2003).

者，而人类则更多地扮演监督者和评估者的角色。

3.GAI、AI- Agent、AGI

从技术发展角度，可以将 AI 分为 GAI、AI-Agent 和 AGI。生成式 AI（Generative Artificial Intelligence, GAI）实践中是指一类基于机器学习和深度学习算法的 AI 系统，其核心能力在于能够根据训练数据集的模式和规律自主创建全新的内容。传统的 AI 是一个“数据－建模－运算－评估”的研究过程。随着深度神经网络和生成式对抗网络(GAN)的发展，机器逐渐可以进行一些内容的创造设计。由 “A 联想到 B，由 A 引起 B，以及由 A 推出 B”的思考方式应该为想象型的智能机器所把握。要在 GAN 的基础上构建想象网络，通过插入式生成、交互生成、相似样本生成等方式，使机器通过想象网络能够具有迁移功能。[①]在 GAI 的认知范式下，每一个事件单元，无论是社会事件、科技进展还是文化现象，都成为 GAI 进行知识生成原料，知识生成逻辑由问题导向转变为事件导向。这种转变意味着知识生成已经从一种基于目的论和解决论的范式，转向了一种基于生成性和可能性的范式。GAI 关键特征在于其创新性输出——包括但不限于文本、图像、音频、代码甚至三维模型等多元形式的数据。[②]

AI Agent 既大家耳熟能详的“智能体”。产业实践将 AI Agent 描述为“集合了记忆(memory)、规划(planning)、执行(execution)和总结(reflection)能力的系统，模拟人类完成某项任务。”[③]大型语言模型（LLMs）和视觉语言模型（VLMs）是构建 Agent AI 系统的基础。LLMs 擅长任务规划和蕴含丰富的世界知识，而 VLMs 如 CLIP 提供了与语言对齐的通用视觉编码器和零样本视觉识别能力。这些模型使得 Agent AI 系统能够理解和处理文本或视觉输入，为开发能够理解和响应多模态输入的 AI Agent 提供了可能。Agent AI 分类为通用代理领域、具体代理、行动代理、交互代理、模拟和环境代理、生成代理、AR/VR/混合现实代理、知识和逻辑推理代理、情感推理代理、神经符号代理、LLMs 和 VLMs 代理等。[④] Agent AI 采用多模态系统算法（MAA）系统。MAA 系统是基于对多模态感官输入的理解，在给定环境中生成有效行动的算法系统。随着大型语言模型（LLMs）和视觉语言模型（VLMs）的出现，提出了许多 MAA 系统，这些系统在从基础研究到具体领域皆有应用。MAA 系统通过整合各自领域的传统技术迅速发展，它们共享共同的兴趣，如数据收集、基准测试和伦理观点，突破经典算法的物理局限，属于具有潜力的“跨界”算法。

[①] 王彦杰：《机器想象何以可能? ——基于发现的视角》，载《自然辩证法研究》2022 年第 9 期，第 36 页。

[②] 米加宁、董昌其：《大模型时代：知识的生成式“涌现”》，载《学海》2024 年第 1 期，第 91 页。

[③] 腾讯研究院：AI Agent 的终极未来，2025 年 7 月 30 日，https://news.qq.com/rain/a/20250730A075RS00

[④] Durante Z, Huang Q, Wake N, et al. Agent ai: Surveying the horizons of multimodalinteraction[J]. arXiv preprint arXiv:2401.03568, 2024.

AGI 是 AI 的最高级别的理想目标。AGI 应具有和人类持平或者超越人类的智能，可以在不限于领域推广自身智能，能够进行反思，甚至具有根本性的洞察力和创造力。[①]理想的 AGI 系统应构建起透明化架构和可追溯的决策流程，具有自我监督及自我调节功能，能够积极阐明决策逻辑，还能真诚地评价自身存在的风险与局限，明确表述认知边界的所在，进而提升系统的可靠性和信任度。未来的 AGI 系统要在架构设计上实现机器自省，需具备三个核心机制：一是配备元认知监测机制，嵌入系统结构中以实时监控各个认知模块的表现，通过对异常模式识别、决策路径的剖析以及计算偏差的发现与校正，实现全面的自我监管与调整；二是建立多视角融合机制，确保认知理解的稳健性，这意味着模型需要有能力协调并综合来自不同视角的理解，力求形成统一而全面的世界观；三是主动设定认知边界，确保世界模型的可靠性，系统应能精确评估自身认知能力的局限，并通过与人类用户的主动交流，诚实地反映认知局限性及潜在风险，接纳人类反馈以指导改进，形成“人机协同解释”机制，规避超越能力边界所带来的高风险。[②]根据现有预期，AGI 时代的到来将会是必然的，那么，现有法律框架对规制高风险 AI 就显得尤为必要， AGI 在技术上是与 GAI 和 Agent AI 一脉相承的。

（二）高风险 AI 的表现形式

可以从两个角度理解 AI 的高风险表现形式，一是具有身份特征的高风险 AI,高风险代表着 AI 的身份标签。二是从后果角度去理解 AI 造成的高风险。两种维度具有交叉性，被冠以高风险身份标签的 AI 不一定会真的造成高风险后果，而造成高风险后果的 AI 并不一定会被冠以高风险 AI 的身份标签。在高风险 AI 的法律治理研究中，可暂时将两种维度一并理解。学界和实务界通常将 AI 的高风险表现形式分为伦理性高风险和技术性高风险。伦理性高风险不考虑 AI 的技术特征。

1.AI 伦理高风险

1.1AI 操纵高风险

顾名思义，“操纵类”AI 是指具有操纵功能的 AI。如果 AI 不被视为具有自主意识的法律主体，那么“操纵类”AI 的关键点就在于：1.操纵的法律意义。2.“利用 AI 进行操纵或者部署 AI 自主操纵的行为”的法律意义。3.操纵类 AI 对人们造成的确切危害。

域内外关于操纵类 AI 法律定义其实并不明确。即使美国和欧盟明文规定应该禁止“操纵类”AI，但却并没有对操纵类 AI 作出过多的解释。[③]也有学者指

[①] Franklin, S. (2007). A foundational architecture for artificial general intelligence. Retrieved from https://digitalcommons.memphis.edu/ccrg_papers/86

[②] 任天知、沈浩：《从 Sora 到“世界模拟”：视频大模型的技术原理、应用场景与未来进路》，载《新闻爱好者》第 2024 年第 6 期，第 31 页。

[③] 美国的一些州颁布了保护未成年人的规范，里面明确禁止网络服务提供者禁止操纵儿童，如《加州适龄

出，由于操纵并不像传统行为那样可以析出确切的组成部分，所以并不可能向传统概念定义方法那样去囊括其概念，相较而言，采用“隐喻+原始属性概念”（prototype theory of concepts）”的方式可以剖析操纵的具体概念，比如“提线木偶”，操纵者就是提线木偶的游戏者，而被操作者就是提线木偶。而原始属性概念其实是一种组合或者集群概念。有些事物群体并不会呈现出每个事物都具有的共同特征，这些事物更像一个复杂、相似、重叠和纵横交错的网络，这些事物有时是整体相似，有时是细节相似，这写事物的相似性类似于“家族相似性”，使家族成员彼此相似的不是所有人共有的一组特征，而是重叠的相似性，即出现在不同成员身上的一组给定特征的不同组合，以原始属性概念的方式将它们组合是自然的。操纵行为就具有这样的“家族相似性”。由此，操纵的法律定义可为：操纵是一种独特的力量，操纵者能将这中力量将非强制的将自身意志强加于人，强加方式有时会避或颠覆理性规则，有时却不用如此，操纵者往往需要洞悉并被操纵的决策、行动或欲望，也经常营造和构造易于操纵的环境，在不剥夺被操作者的能动性的情况下，又能常驾驭被操纵者的意志，操纵行为后果为既可以推进操纵者的目标，又可能促进被操纵者目标，常以完全透明的方式进行。”[①]

1.2 AI 失控高风险

简明扼要地讲，“失控类”AI（AI unbound）泛指具有高度自主性摆脱了人类控制的 AI。自 AI 技术诞生以来，就有人们就有对 AI 失控担心。AI 失控会导致严重后果，和 AI 技术特征并不具有本质关系，并不是尚未到来的 AGI 才会发生失控，对于正在发展和利用 AI 也有失控的风险。AI 失控并不是隐喻或者预言，是法律亟需面对的问题。法律处理 AI 失控无外两种方式，其一，预防失控的发生，其二，填补，复原和挽救失控发生后的严重后果。如果“失控类”AI 造成灾难的严重后果，就应该被禁用。

截止目前，学界和实务界对“AI 失控”并没有明确的法律定义，暂且可以通过技术层面来分析，衡量“失控”的要素为：其一，AI 技术运行过程会不可避免的发生“不可预测”或者“意料之外”的结果。AI 的技术机理为机器算法，高阶算法的算法运行者本身无法控制算法结果，如若算法结果错误时，通过算法补丁来修复算法，又加上，算法研发工作是群体性质的，所以补丁更新可由不同人进行，很多时候，补丁算法的人和制造算法漏洞的人往往并不是同一个人，以至于算法设计者最终无法“窥探”精确的算法运行流程，所以，算法失控时有发生。《生成式

儿童规范》。

[①] Shlomo Cohen，The Concept And Ethics Of Manipulation，Cambridge University Press，2025，pp.50-65.

人工智能服务管理暂行办法》第 4 条第 5 项规定， 提供和使用生成式人工智能服务，应当遵守法律、行政法规，尊重社会公德和伦理道德，基于服务类型特点，采取有效措施，提升生成式人工智能服务的透明度，提高生成内容的准确性和可靠性。由此条可以看出，如若生成的内容不准确（失控）只要在法律允许范围内是可以容忍的。

技术上，AI 失控的主要原因是背离人们的目标，业界常用“人机未对齐”来描述 AI 失控。[①]AI 失控本质是对人类信任的背离。无论 AI 技术是否在客观上具有高风险，如何构建 AI 和人类之间的信任始终都是 AI 理论的核心议题，而盲目信任本身意味着“可能被背叛”的高风险。人们对于失控类 AI 的正确态度应是“零信任”。[②]遗憾的是，失控类 AI 会以人们通常注意不了的方式背离人们的目标。为了防止 AI 失控，人们能够采取的常见方式为限制 AI 与外界沟通和施加影响能力。限制方式通常面临双重挑战：一方面是技术难题——AI 自身须维持和提供有效信息的能力确保其不会逃逸。另一方面涉及 AI 运行背景问题，AI 运行者可能处于各种动机放松对 AI 的限制。该双重挑战极易容易被攻破，足够智能的 AI 会以某种方式找逃逸的方法：人类有意的分享自主权力给 AI，且在 AI 达到自主权力分享条件时，还不断给与 AI 技术人员激励以促成更加自主的 AI——扩展现有 AI 系统的控制范围或者升级现有 AI 系统或新增具备前所未见能力的新 AI 系统，若潜在利润足够丰厚，部分企业仍倾向于投资将业务置于自主人工智能的掌控之下，这些都是故障人工智能可能引发重大风险的隐患。2.当人类具有犯罪意图或者恐怖主义目的等本身违背人类伦理的目标时，AI 的行为和人类都属于失控状态。3.AI 将目标理解成“自己应享有不受外部约束自主行动的权利的意识主体”。4.AI 运行者基于对 AI 释放环境安全性误判等等。[③]

1.3“伤害个人利益”的 AI

如果 AI 操纵高风险应被法律谴责的重点为行为模式，AI 失控高风险应被法律谴责的重点为 AI 对人类主体地位的挑战，那么“伤害类”AI 应被法律谴责的是重点当然为对人们带来伤害的结果。AI 伤害性高风险是泛指 AI 对人的生命安全造成严重威胁或伤害。[④]

高风险 AI 广泛伤害存在于各种法律宪章和自然享有的人格权利：1.生命健康权，自动驾驶汽车对人们带来的伤害。2.隐私权和数据保护权。AI 会分析会刺探人们的数据，对人们进行监控。3.声音权。声音作为关乎自然人人格的重

[①] 吴映平：《生成式人工智能的创造性失控与技术拟人化伦理秩序的重构》，载《 武汉大学学报(哲学社会科学版)》2025 年第 5 期，第 31-41 页。
[②] 吴冠军：《大语言模型的信任问题与资本逻辑》，载《当代世界与社会主义》2023 年第 5 期，第 12 页。
[③] Eurasia Group's Top Risks For 2025.
[④] Lorin Brennan, "AI Ethical Compliance Is Undecidable," Hastings Science and Technology Law Journal 14, no. 2 ，2023，pp: 311-338.

要人格权益，AI 技术服务开发者在使用自然人声音时应当起到较高的注意义务。[①]4.公平审判权，比如威斯康星州法院使用了 COMPAS 风险分析系统使受决策者承担了过重的刑法负担。[②]5.行动自由权，防止被 AI 操控和限制。[③]6.平等权。AI 并非只满足于扮演知识生产参与者、行动者的角色，它背后固有的权力扩张、资本冲动，它所拥有的大数据和超强算力，注定人与机器在数字化知识生产过程中无法“平等化构建”。[④]

1.4.伤害公共利益的 AI

影响公共利益的高风险 AI 包括但不限于造成以下风险：1.AI 的安全风险。“AI 的安全风险一定程度上来源于 AI 内部，同时驱动 AI 不断进行价值观对齐。但是诸如价值观本身都是有政治安全风险、伦理安全风险、意识形态安全风险的，那么 AI 就不可能对齐。”[⑤]2.AI 的道德风险。Sora 文生视频的“黑箱效应”使 AI 的道德和法律归属变得复杂，对社会信任和接受度构成一定程度的消解。[⑥]3.AI 的技术失效。[⑦]以 ChatGPT 为例，这些技术风险主要源于生成性内在缺陷和训练数据时效性的制约。[⑧]4.AI 的信息滥用风险。超大模型和大模型都可能生成不真实的信息，而如果完成消除这种幻觉信息，将会抹杀 AI 的创新，如果滥用信息将会产生恶劣影响。[⑨]5.涌现风险。涌现的不确定性可能为具体应用领域带来知识偏差、模型鲁棒性、偏见与歧视，以及责任方面的挑战。[⑩]6.AI 产业风险，其包括 AI 供应链安全，AI 安全性评估、AI 生成内容安全。

2. 技术层面的高风险

由于技术层面而引发的高风险为 AI 技术高风险。就风险后果来说，AI 技术高风险和 AI 伦理高风险有交叉和融合之处，并不是两种完全平行的分类。AI 伦理高风险很大程度上也由 AI 技术导致。不过，AI 技术高风险旨在通过 AI 的技术逻辑分析其风险类型。AI 技术类型虽然复杂，但是技术基础架构还比较稳定，

---

① 北京互联网法院课题组：《AI 生成声音侵害声音权益的法律认定——以殷某某诉北京某智能科技公司等人格权侵权案为例》，载《法律适用》2024 年第 5 期，第 133 页。

② 黄国栋：《比较法视野下智慧法院建设的中国经验、 实践困境与路径优化》，载《法律适用》２０２３年第 3 期，第 130 页。

③ Access now organization：Human Rights In The Age Of Artificial Intelligence

④ 曾白凌：《大模型、大数据、新精英：数字技术对知识生产的征服与全球性重构》，载《现代出版》2023 年第 6 期，第 33 页。

⑤ 郭全中、张金熠：《AI 向善：AI 大模型价值观对齐的内容与实践》，载《新闻爱好者》2023 年第 11 期，第 21 页。

⑥ 令小雄、王鼎民、唐铭悦：《ChatGPT 到 Sora：Sora 文生视频大模型对影视创作的机遇、风险及矫治》，载《新疆师范大学学报（哲学社会科学版）》第 2024 年第 6 期，第 134 页。

⑦ 童祁、胡晓萌：《不要温和地走进与 AI 共生的时代 ——追问大模型时代的人机关系》，载《上海文化》2023 年第 8 期，第 74 页。

⑧ 秦策：《操作性智能标准下的 GPT 参与司法》，载《中国应用法学》2023 年第 4 期，第 103 页。

⑨ 高奇琦：《从大国协调到全球性机制：人工智能大模型全球治理路径探析》，载《当代世界》2024 年第 6 期，第 13 页。

⑩ 顾东晓、赵旺、王晓玉、汤志伟、朱凯旋：《大模型“涌现”视域下的智能健康医疗知识管理体系变革与创新发展》，载《图书情报知识》2024 年 9 月，第 5 页。

业界通常将 AI 分为通用型 AI 和具体应用 AI，以下详述之：

2.1 通用性 AI 的高风险

通用型 AI (General-Purpose AI Models)—— “AI ACT”将其称为 GPAI,我国的《人工智能示范法》称其为基础模型研发者，为了行文简便，下文将通用型 AI 也简称为 GPAI。“AI ACT”对 GPAI 的描述为，GPAI 无论以何种以何种方式投放市场都显示出显著的通用性，能够胜任各种不同的任务，并可集成到各种下游系统或应用中。易言之，人们可以把通用型 AI 理解为土壤，这些土壤可以设计、塑造、加工、组件、构造成不同的建筑物（具体应用类 AI）[①]。如何在技术上定义 GPAI，业界并无统一答案，我更无明确的法律定义，但是欧盟给出了一个明确的技术标准——即当模型的训练计算量超过 10^{23}浮点运算次数/秒（FLOP），且能够生成语言、文本转图像或文本转视频时，该模型将被视为 GPAI 模型。[②]浮点运算（FLOPS——Floating-point Operations Per Second）是计算机的衡量算力的基本计算单位，直接衡量处理器（尤其是 GPU 和科学计算 CPU）进行科学计算、3D 图形渲染、AI 训练（大量使用矩阵乘法，本质是浮点运算）等需要高精度计算的能力。事实上，“AI ACT”给出的衡量标准确有疑问，为什么不将其他计算单位纳入计算阈值？比如整数算力（OPS——Operations Per Second）/ TOPS——Tera Operations Per Second）。整数算力为直接衡量处理器执行逻辑运算、整数加减、数据移动、控制流程等操作的能力，在 AI 推理（Inference）、网络数据处理、通用计算等领域亦非常重要。

由于 GPAI 的高风险主要体现在两个方面：

1.系统性风险。GPAI 的系统性风险是指 GPAI 技术系统本身发生的多层次、多维度风险。首先是交互层风险用户通过多模态交互接口（文本、语音、视觉、图形）向 AI 智能体系统提供输入。风险发生点为利用提示词注入、信息泄露、绕过、建构违规输出、侧信道注入、篡改/欺骗等手段，直接影响用户与智能体系统的交互。这些攻击可能发生在输入处理阶段，试图改变智能体的意图或窃取敏感信息。其次，为 Agent 集群层风险，该层是 AI 智能体系统的核心处理单元，包含主 Agent 及其模型层 (Model Layer)（LLM，负责推理、理解、生成、规划）和认知层（负责认知与意图管理、记忆与状态管理）。此外，还包括工具交互层（工具选择与调用逻辑）和环境交互层（感知与观察、Agent 通信与协作）。模型层主要面临幻觉、对齐不足、对抗样本、模型规避等风险，直接影响 LLM 的输出质量和可靠性。认知层主要功能为学习记忆功能、认知和意图管理功能，主要面临记忆投毒、篡改/注入的风险，可能导致智能体的意图劫持、欺骗/误导和

① Novelli, C., Casolari, F., Rotolo, A. et al. Taking AI risks seriously: a new assessment model for the AI Act. AI & SOCIETY,Vol.39,2024,pp. 2493–2497.

② Guidelines on the scope of the obligations for general-purpose AI models.

决策干扰/注入等记忆风险，或则核心逻辑能力恶意利用风险。工具交互层的主要风险来自不当的工具选择与调用逻辑，可能导致智能体误用或滥用工具，执行高风险操作。环境交互层的主要作用在感知与观察，可能面对通信投毒、篡改/注入等风险，可能导致智能体从环境接收到虚假或被操纵的信息。最后为部署层。GPAI 的部署主要为本地 MCP 服务和远程 MCP 服务。它们分别通过 STDIO 和 SSE 协议提供工具、提示词和资源的访问，以及可选的知识库、API 和代码执行 (Code Execution)功能。MCP 主要面临提示词注入 (Prompt Injection)、信息泄露 (Information Leakage)、本地越权访问/敏感信息窃取 (Local Privilege Escalation/Sensitive Information Theft)、命令注入/代码执行 (Command Injection/Code Execution)等风险。这些风险通常发生在 MCP 服务的本地部署环境中。网络 MCP 服务：MCP 支持以 http/SSE 协议为传输层的网络 MCO 服务，除了与本地 MCP 服务相似的风险外，还面临更复杂的网络攻击风险，如跨智能体/信息越权访问 (Cross-Agent/Information Privilege Escalation)，以及身份冒充/越权访问 (Identity Spoofing/Privilege Escalation)和流氓智能体 (Rogue Agent)/决策操纵 (Decision Manipulation)等，这在多 Agent 通信中尤为突出。概而论之，GPAI 的系统性风险是通过复杂的、动态的传导路径相互影响，对整个系统的安全性、可靠性和合规性构成威胁：（1）交互层到 Agent 集群路径。外部输入（如恶意提示词注入）可以直接攻击 LLM 核心，或影响感知模块，进而传导到认知层和工具交互层。（2）Agent 内部模块传导路径。GPAI 的输出问题（如幻觉、对齐不足）会传导到认知层，影响意图管理、记忆和规划，最终导致错误的工具选择或行动。(3) Agent 集群到 MCP 服务路径：智能体（如工具调用逻辑）发出的指令会通过 MCP 协议传递给本地或远程 MCP 服务，若这些服务存在漏洞，则可能被利用进行攻击（如命令注入）。（4）MCP 服务到 Agent 集群路径：MCP 服务提供的被污染的提示词或资源会回传给 Agent 集群，进一步影响 LLM 推理、记忆和规划。（5）Agent 间/与其他 Agent 通信：智能体之间或与其他 Agent 的通信如果受到投毒或身份冒充，可能导致错误信息在多 Agent 系统中传播，或流氓智能体对其他 Agent 的操纵。

2.“毒树之果”的风险。无论前述风险无论发生在何种范围，都会毫无保留的传递到下游应用，并且只会有放大的风险，例如，不当输出处理类漏洞、信息泄露类漏洞、提示注入类漏洞、缺乏针对无限制消耗类攻击的防护措施、传统安全漏洞等普遍风险都会发生传导。[①]还需特别强调的是，GPAI 的开源会加速其“毒素之果”的风险。“开源”在 AI 技术亦不是具有明确定义的术语，开源是指 AI 的

[①] https://www.news.cn/tech/20250916/f42c8fe3530e42b6952d60d997b1a8c0/c.html，最后访问时间 2025 年 9 月 25 日。

开放性。开源应是光谱化的，而不应该只有开源或者闭源的区别，与开源代码不同，AI 由一系列组件组成，每个组件都可以由开发者组织保留或沿着上述开放性范围发布。首先，开源意味着包括黑客在内的恶意行为者拥有与善意贡献者相同的访问权限。他们可以详细研究代码，搜索可利用的漏洞，包括已知漏洞和安全漏洞。其次，开源依赖于部署环境，开源虽可以减少不同 AI 开发人员之间的能力差距来促进 AI 力量的去中心化，但这也让 GPAI 模型处于更加不确定性的状态。最后，开源会让 GPAI 原有的运行者失去一部分控制权，从而加速“毒树之果”的传播。[①]

2.2 具体应用 AI 的高风险

具体应用 AI 主要指 B 端（商业端）或者 C 端（消费者端）的应用。具体应用 AI 技术发展如火如荼，种类繁多，由于其用途不一致，并不是所有具体应用 AI 都会导致高风险，现选取学界和实务界认为高风险多发领域的具体应用 AI 的典型范例，分析如下：

2.2.1 金融证券 AI 的高风险

常见的金融证券 AI 有：智能投顾和高频交易 AI。智能投顾是指没有或较少有自然人投资顾问的直接参与，而是以大数据为基础，通过算法自动化提供投资者教育、客户洞察、投资组合分析与选择、资产配置、交易执行、投资组合再平衡、税务规划、遗产传承等一项或多项金融理财建议与执行的服务。[②]高频交易 AI(程序化交易 AI)是指通过计算机程序自动生成或下达交易指令，在证券交易所进行证券交易的行为。[③]智能投稿 AI 高风险表现形式为：1. 投资决策错误风险。2. 扰乱金融市场风险。3.全权代理的风险。4.投资范围违规风险。5.网络安全风险。如数据安全和个人信息保护风险，APP 安全认证风险。[④]

2.2.2 医疗诊疗 AI 的高风险

顾名思义，医疗诊疗 AI 是指运用在医疗领域的 AI。由于医疗 AI 的技术机理主要是模式识别和词汇共现概率预测，而非基于医学病理、生理机制的理解能力，而是通过对语言共现的模式识别来“生成”答案，AI 的技术局限性体现在：其一，AI 缺乏真实医学推理能力。其二，AI 在医学推理中欠缺因果推理能力、个体化诊疗能力、与多源医学信息整合能力。[⑤]其三，医疗数据质量与代表性并不

[①] UK:Open Source Software Best Practices and Supply Chain Risk Management.
[②] 朱紫涵、余涛，《智能投顾的国际经验及借鉴》，载《东南大学学报(哲学社会科学版)》2021 年第 1 期。
[③] 王策、刘蓓：《智能投顾规制的理论反思与制度构建》，载《吉林大学社会科学学报》2025 年第 3 期，第 201—210 页。
[④] 张爽：《金融领域生成式 AI 技术的应用、风险与应对》，载《债券》2025 年第 5 期，第 69-75 页。
[⑤] Akhil Vaid：Medical AI can transform medicine — but only if we carefully track the data it touches，Nature ,vol,682,2025,pp864-866.

能如此，具有局限性和偏差性。其四，AI 的事实校验能力不足。[①]基于技术缺陷，医疗 AI 表现的高风险为：1.AI 误诊和错误推荐治疗方案。2.AI 虚构医学研究与数据来源。3.AI 在影像分析中的错误识别。4.药物剂量建议错误。5.加速人类医生的医疗机能退化。[②]

2.2.3 公共部门 AI 的高风险

AI 嵌入政务办公、城市治理、民生服务及辅助决策等公共部门治理，即为公共部门 AI。[③] “AI ACT”直接把涉及移民、就业、执法领域的 AI 视为高风险 AI。公共部门 AI 高风险并不只是伤害公共利益，也会伤害私人利益。[④]根据现有研究，公共部门 AI 的高风险体现在：1. 政务工作中涉及的敏感数据，核心数据，重要数据，公民个人数据、国家秘密数据的隐私和安全风险。2.对公民的不平等、歧视风险。[⑤]3.公共部门的信息准确性与权威性风险。4.合规性与伦理风险。5.政务人员的技术依赖风险等等。[⑥]

二、高风险 AI 的正面效益

高风险 AI 接纳技术的发展更新，战略性地利用 AI 驱动的工具，解决道德、隐私和安全问题，培育适应性和创新性的文化，发展新的技能和能力，投资于技能发展，培育创新文化，解决伦理道德、隐私和安全问题，参与政策制定和宣传，接纳合作和跨学科的实践。[⑦]高风险 AI 的正面效益源自于 AI 的正面效益。学界比较关注 AI 带来的风险和危害，但是对 AI 所带来的正面效应研究较少。此番情景这也许来自于人们在 AI 技术革新带来的冲击力总是认为法律总应照顾弱势利益的制度偏好。AI 在党中央的高层设计中是被赋能时代发展的，人们均应能享受 AI 技术带来的时代福利。[⑧]

（一）人机交互效应

[①] 李润生：《人机协作的法律形态及过失认定——人类医生与医疗 AI 的共处之道》，载《北京航空航天大学学报(社会科学版)》 [2025-09-25]，第 1—16 页。

[②] Alber, D.A., Yang, Z., Alyakin, A. et al. Medical large language models are vulnerable to data-poisoning attacks, Nature Medicine vol 31,2025, pp. 618–626.

[③] 李姝婷、邱茜:《当 AI 成为“公仆”:政务服务数字人类型对政府信任的影响研究》，载《公共行政评论 》2025 年第 4 期，第 1—21 页。

[④] 邱茜、李姝婷：《技术失灵还是互动缺失？基于情境实验的 AI 政务服务失败对公众满意度影响研究》，载《电子政务》[2025-09-25].第 1-13 页。

[⑤] 李立清、栾文婷：《生成式人工智能嵌入公共服务治理的作用机理与风险防范》，载《南昌大学学报(人文社会科学版) 》2025 年第 3 期，第 113—124 页。

[⑥] 吴建雄、宋阳：《人工智能辅助纪检监察办案的法律风险及其控制》，载《广州大学学报(社会科学版) 》2025 年第 4 期，第 118-131 页。

[⑦] ［美］丹尼尔·D·李著，管斌、宋博文译《法律与人工智能：用ＣｈａｔＧＰＴ塑造法律实践的未来》，载《财经法学》2023 年第 4 期，第 192 页。

[⑧] 孙会岩：《ＣｈａｔＧＰＴ 技术视野下政治秩序重构及其风险治理》，载《学术界》2023 年第 8 期，第 89 页。

AI 具有理解的迅速性和高效率性，形成理解的知识背景的广泛性，理解的规范性等优势。”[①]当 AI 与用户进行对话时，它可以通过自然语言处理技术来解析理解用户的文本或者语音输入，同时利用视觉技术来分析用户的面部表情和姿态，以更好地理解用户的意图和情感，而在回复用户时，它可以同时生成文字回复和语音合成，还可以通过图像、视频的检索、生成等技术，以多种方式表达自己的输出。这种 AI 的输入输出方式可以更好地满足用户的需求，提供更自然丰富和具有沉浸感的交互体验。[②]具身智能技术尤其适用于公共安全和救援任务，其设备能够在危险环境中快速部署并高效执行搜救任务，完成监测、清理和运输未知危险物质的任务。这些任务对人员健康和环境有潜在威胁，而具身智能的应用大大降低了这种风险。[③]

（二）行业应用效应

多模态预训练模型可以在政务服务领域推进政务办事精准指引和高效审批，在医疗领域，提供专家级的科学诊断建议和提高医疗资源利用效率。在科学研究领域缩短科研实验周期。在金融领域提高金融机构对投资项目、贷款申请、交易风险等方面的分析能力。[④]GPT 辅以监控指定金融产品、交易环节、交易双方用户以及其他风险，及时对风险进行报警和管控。[⑤]在自动驾驶领域辅助提高车载自动驾驶模型的泛化能力；在城市治理领域，为城市治理决策提供全面综合的技术支撑。[⑥]在 AI 技术可以融入主流媒体 + 商业媒体，将其内容导向与质量关乎关键信息在社会层面的健康供给与和谐流动。[⑦]增强人机协作效能未来的教育场景更倾向于从当前的单人机交互走向多人多机交互，逐步进入多智能体辅助教学的时代。[⑧]多智能体系统能整合不同算法和数据源的优势，更加精确地模拟复杂问题解决中的真实协作机制。”[⑨]

（三）作为规制工具的 AI

AI 即可以被治理的对象也可以成为治理对象的工具，比如就混淆迷惑治理

---

① 肖峰：《大模型的理解力之争与理解观新叙事》，载《社会科学》 2024 年第 1 期，第 51 页。

② 刘静、郭龙腾：《ＧＰＴ－４对多模态大模型在多模态理解、生成、交互上的启发》，载《中国科学基金》第 2023 年，第 5 期，第 800 页。

③ 赵杨、张佳怡：《基于具身智能的人智交互体验研究：理论、应用与展望》，载《情报理论与实践》网络首发：2024 年 9 月 26 日，

④ Brightwood, Seraphina & Jame, Henry. (2024). Data privacy, security, and ethical considerations in AI-powered finance.

⑤ 杨望、徐慧琳、王钰淇：《构建 GPT 大模型的经济循环体系：理论框架与发展路径》，载《新疆师范大学学报（哲学社会科学版）》2024 年第 3 期，第 79 页。

⑥ 李耕、王梓烁、何相腾、彭宇新：《CHATGPT 到多模态大模型：现状与未来》,载《中国科学基金》2023 年第 5 期，第 730 页。

⑦ 匡野：《大模型时代主流媒体内容平台建设三重逻辑》，载《中国编辑》2024 年第 4 期，第 41 页。

⑧ Holmes, Wayne. "Artificial intelligence in education." Encyclopedia of education and information technologies. Cham: Springer International Publishing, 2020. 88-103.

⑨ 翟雪松、张丽洁 、 夏亮亮 、 徐鑫、朱强：《基于 GAI 的逆向工程教学思维在人机协作中的应用研究——以编程教育为例》，载《课程与教学》2024 年第 9 期，第 66 页。

主体而言，由于规避者从事的亦为直接的违法违规行为，仅仅利用信息洪流进行掩饰，故而同样不存在正当化理由。通过人工审查辅助与机器学习技术，提高数据分析与智能识别的精度，避免 AI 被不智能的简单处理所蒙蔽。”[①] “技术代码具有伦理意蕴与规范属性。技术代码与伦理规范一样涉及价值的形成与沉淀，其解构与重构过程就是价值生成与沉淀的过程。AI 等高新技术的发展是曲折而复杂、难以被全面把控的,以不变应万变的传统思路未必持续有效,一种更值得提倡的态度是因势而变、以变制变,就有了 AI 对抗 AI 的实际需求。[②]

三、高风险 AI 的法律判定机制

基于风险的动态性及不可预测性，主张将所有的 AI 系统都纳入高风险会导致 AI 运行者承担严苛的监管要求和法律责任而极大减缓 AI 技术和产业发展。在法律程序上缺乏判定机制为高风险 AI 是风险规制路径不得不面临的难题。[③]有学者抨击欧盟分类分级标准的有效性，并且总结出了适应性强的风险判别方案，将 AI 风险的分类方式从分离“输入层＋输出层”的视角，将输入层风险分为已知输入和未止输入两种类型，而输出层分为可预见性输出和不可预见输出两种类型。[④]判定高风险 AI 的方式对于风险治理思路是必不可少的。

锚定 AI 的关键性技术特征之后，厘清高风险 AI 的判定机制就需将焦点集中在何为“高风险”上。在学界，我国专家贡献的《人工智能示范法》（法大版）中对高风险 AI 的表述为“关键人工智能”，其判定规则较为简单：1.应用于关键信息基础设施的人工智能；2.对个人的生命、自由、人格尊严等人身权益有重大影响的人工智能；3.在算力、参数、使用规模等方面达到一定级别的基础模型；4.法律、行政法规规定的其他情形。域外已有法案明确界定了“何为高风险 AI”。首当其冲的应为 2021 年 4 月 21 日欧盟委员会公布的《人工智能法》（以下简称“AI ACT”）其将 AI 风险分为四个等级：不可接受的风险、高风险、有限风险、最小风险。整体上，高风险 AI 确定规则由 “阐述性规则+补充性用例（以附件补充的方式）+调整性规则+排除性规则”构成，较为复杂。[⑤]

---

① 邱遥堃：《法社会学视角下的算法规避及其规制》，载《法学家》2023 年第 3 期，第 72 页。

② 王晓梅、何丽：《人工智能伦理规范建构与芬伯格的技术代码方案》，载《伦理学研究》2022 年第 2 期，第 92 页。

③ 秦瑞标:大模型时代生成式人工智能治理的逻辑更新与回应路径———以《生成式人工智能服务管理暂行办法》为视角》，载《时代法学》2024 年第 3 期，第 33 页。

④ Keri Grieman & Joseph Early, A Risk-Based Approach to AI Regulation: System Categorisation and Explainable AI Practices, 20 SCRIPTed 56 (2023).

⑤ 其他法域的高风险判定规则为：2025 年 2 月 25 日，美国的弗吉尼亚州颁布了《高风险人工智能研发者和部署者法案》（the High-Risk Artificial Intelligence Developer and Deployer Act）（以下简称弗吉利亚州法案）。该法案将高风险 AI 定义为：“任何有目的自主或者作为实质性因素作出重大决策（consequential decision）的 AI 系统。何为“consequential decision” ,该法案的解释为，为任何消费者（仅在个人或家庭环境中活动的自然人，排除了商业或就业环境中行事的自然人）提供或拒绝具有实质法律意义或类似影响的

遗憾的是，高风险 AI 的判定机制并不明朗，主要表现为：首先，判定对象复杂。风险的高低和 AI 技术复杂程度并不呈现正相关。在技术层面较为简单的深度合成依旧会带来高风险，而技术层面较为复杂的 AI Agent 如若控制得当，不必然带来高风险。高风险 AI 在复杂的技术架构的嵌套中难以确定合适的判定切入点。针对复杂技术实践划定的伦理规范和监管实践对判定高风险 AI 并不具有实际意义。[①]

其次，判定规则模糊。我国学者建议稿中的高风险 AI 判定规则较为简单，尚需进一步解释和探讨。而纵观域外高风险 AI 判定规则，欧盟，美国，澳大利亚，加拿大都无一例外的采用了将不确定性的高风险 AI 概念通过判定框架逐渐确定，以供法律政策制定者和 AI 运行主体参考到判定方式，然而看似完整的判定逻辑实则具有局限，具体体现为：1.阐述性规则措辞抽象。2.补充性用例容易导致解释学循环困境。3. 阐述性规则具有不确定性。4.排除性规则存在留白等问题。[②]

最后，判定权力集中。具有判定 AI 是否为高风险的法律权力主体为判定主体。域外法案都不约而同将判定权力赋予了立法机构。“AI ACT”将判定高风险 AI 权力赋予了欧盟委员会，而美国，澳大利亚，和加拿大相关 AI 法案也将判定权力赋予各自国会或者州议会。[③]此举问题在于，立法机构是否足够了解发展得如火如荼 AI 技术，且能做出客观科学判断？事实并非如此。各国相关 AI 法案颁布进程并不一帆风顺。[④] “AI ACT”一经颁布就招致多重批评，这些批评焦点在于民众对欧盟立法机构判定能力的不信任上，而我国高风险 AI 的判定机制注意尽力避免此类问题。[⑤]

四、高风险 AI 的治理理论

当促进发展与防范风险通过党的政策和治理实践被确立为法政策目标后，AI 技术发展初期的立法将防范风险作为第一原则。人们也逐渐认识到发展 AI

任何决定，该法案提供了什么不是高风险 AI 的法律阐释和白名单。法律阐述为:1.执行狭隘程序的 AI。2.改进先前人类活动成果的 AI。3.检测自动决策系统及其与先前自动决策相比的偏差的 AI。4.预先评估高风险 AI 的 AI。而白名单上的技术有 18 种，分别为：1.不使用人脸识别的反欺诈技术。2.反恶意软件技术。3.防病毒技术。4. 人工智能视频游戏。等等。2024 年 10 月，澳大利亚政府人权委员会颁布了《高风险人工智能强制性法律边界》（mandatory guardrails for AI in high-risk settings），这些法律框架界定的高风险 AI 为，在没有正当理由的情况下，不利影响澳大利亚人权法承认的个人权利以及澳大利亚的国际人权法义务的 AI。“反歧视”成为了澳大利益高风险 AI 法律规则的主题，而对于美国弗吉利亚州高风险 AI 法案（ High-Risk Artificial Intelligence Developer And Deployer Act）亦然。

① 薛桂波、赵建波：《从“应当”到“是”：人工智能伦理规范实践策略探析》，载《自然辩证法研究》2023 年第 1 期，第 90 页。

② EU Artificial Intelligence Act

③ Schuett, Jonas. "Risk management in the artificial intelligence act." European Journal of Risk Regulation.vol 15.no.2 ,2024,pp 367-385.

④ Concerning Consumer Protections In Interactions With Artificial Intelligence Systems.

⑤ Warthon, M. Restricting access to AI decision-making in the public interest: The justificatory role of proportionality and its balancing factors . Internet Policy Review, vol.13.no.3.2024，pp1-32.

产业的重要性,促进发展便成了一种法政策目标。[①]未来人与 AI 并非对立和竞争的关系，而是 “人机”相互合作关系。人机合作并非仅指人类对于 AI 所持的态度，还意味着社会系统与数字系统的深度融合，这种融合真正意义上重塑了整个复杂社会巨系统的形态和结构。人机合作应当同时考量它对数字系统和线下生活系统的意义和影响，将两个系统结合起来从整体上进行科学的利弊权衡，形成妥当的法律判断和解决方案。”[②]高风险 AI 运行边界分为伦理边界和法律边界。在 AI 时代，随着伦理规范化在学界的研究逐渐深入，两者之间的界限渐渐模糊。[③]高风险 AI 法律治理的理论基础来源广泛，除了法学基础理论，亦有哲学，伦理学以及社会学等其他学科领域的理论。这些理论大都聚焦于伦理范畴，主要影响人们如何评价高风险 AI， “如何构建 AI 和人类之间的信任”是这些 AI 理论的核心议题。[④]高风险 AI 的治理理论奠基与哲学理论和公共社会学理论、公共学理论之上，下面分述之：

（一）高风险 AI 的哲学道德理论

哲学体系非常庞杂，学界讨论得比较激烈理论有：1.新唯物主义理论。在哲学角度，是否应该把 AI 视为“意识”或者“物”引发了争论。从当代欧陆哲学的“新唯物主义”回到马克思本人的新唯物主义。在与人类中心主义的理论“对赌”中,“新唯物主义”押上了物质自我涌现的“法宝”,试图唤醒“非人类行动者”在理论中的积极形象。[⑤]2.具身理论。梅洛·庞蒂的具身理论提供了一套哲学框架，探索性地推演出一组可以作为未来具身智能体发展，实现理想目标的规范性条件。[⑥]这些规范性条件包括人们应该允许 AI 智能体的身体和感受器参与经验的形成，尤其是通过感知外界而生成经验数据,而不是把现成的、加工处理过的高质量数据作为输入项；[⑦]智能体还需要一种身体图式，原初地参与经验的生成。[⑧]具身智能体应能理解外部世界，而且能够根据身体的认知行动，这种行动并不一定是由智能体的中心处理装置发出任务决策，智能体可以不再依赖于调动外部设备来执行某些行为任务。[⑨]3.康德道德哲学。康德道德哲学主要规范 AI 的行动

---

① 张宇帆：《论人工智能政策与法律协同治理》，载《东方法学》2024 年第 5 期，第 194 页。
② 梅夏英：《复杂系统与智能涌现: 未来数字法研究的范式图景》，载《法学家》2024 年第 5 期，第 59 页。
③ Rodrigues, Rowena. "Legal and human rights issues of AI: Gaps, challenges and vulnerabilities." *Journal of Responsible Technology* 4 (2020): 100005.
④ 吴冠军：《大语言模型的信任问题与资本逻辑》，载《当代世界与社会主义》2023 年第 5 期，第 12 页。
⑤ 牛小雪：《当代欧陆哲学 “新唯物主义” 的 “物活力” 想象及其评析》，载《哲学动态》2022 年第 12 期，第 43 页。
⑥ 张昌盛：《从具身智能到具身智能体》，载《北京工业大学学报（社会科学版)》2023 年第 6 期，第 50 页。
⑦ Manzotti, Riccardo. "Embodied AI beyond embodied cognition and enactivism." Philosophies 4.3 (2019): 39.
⑧ 冯子轩（编），《人工智能与法律》，法律出版社 2022 年 11 月版。
⑨ Lungarella, Max. Exploring principles towards a developmental theory of embodied artificial intelligence. Diss. Universität Zürich, 2004.

决策。在 AI 行动决策的依据中，康德道德理论在两个基本前提的情况下可转化为三条基本的人工智能伦理原则：其一，不侵犯原则。[①]任何行动不能侵犯行动者的必要善。其二，平等原则：所有行动者的必要善平等。其三：援助原则。当行动者 A 的必要善处于威胁之中，如果 A 的必要善的维护完全取决于 B 的援助，并且 B 完全有能力援助，那么，在援助行为不会导致任何行动者的必要善受到侵犯的情况下，B 应当援助。[②]4.机体哲学视角。机体哲学视角为如何判断高风险 AI 的伦理边界设定了步骤。首要的分析内容：该个体或群体接收过哪些关于 AI 的符号？AI 可能作为广延性质的物体产生的形态符号或者姿态符号。AI 社会符号如何受到 AI 作为互动个体的影响，分析各个符号的价值，并结合所有的符号与价值得出结论。简单地说，如果利益带来的影响更多，价值即正向；伤害带来的影响更多，价值即负向。该步骤既可以定向也可以定量。[③]

（二）高风险 AI 的社会学理论

社会理论也为搭建高风险 AI 法律框架提供了方法论和思考角度。行动者网络理论运用在学界得到了很大赞同，人类中心主义和后人类中心主义的争议也如火如荼，扎根理论为高风险 AI 的研究提供了质性方法论。

1.行动者网络理论

行动者网络理论是由布鲁诺·拉图尔和其他社会学者在 20 世纪 80 年代提出的一种理论框架，其基本观点是技术和社会是不可分割的，它们通过网络连接在一起，在这一网络中，“行动者”既可以是人也可以非人的实体，如技术、物体、概念等，每一个行动者都在互相影响和塑造网络， AI 等大模型媒介作为一种革命性的技术已经成为了一个最具活力的“行动者”，正在深度卷入我们的社会结构，重新构建我们社会的行动者网络。[④]行动者网络理论不固守传统法律中的主客体两分的研究范式。[⑤]行动者网络理论中的关键行动者形成了 AI 伦理规则的执行和走向，分别是 AI 模型研发者，数据分析公司，AI 服务提供商和 AI 社交平台。关键行动者会影响其他行动者的理念和行动，也是风险规制环境的重点环节。[⑥]

2.人类中心主义和后人类中心主义理论

人类中心主义理论认为，自主性是人能否享有尊严的首要条件，一旦丧失

[①] 李亚明：《当代道德规范性研究与人工智能伦理设计》，载《哲学动态》2023 年第 5 期，第 111 页。
[②] 第一条原理：我们应当就每一个行为向行为所涉及的行动者证明，该行为是由没有理由拒绝的原则所支持的；第二条原理：行动者不得不欲求必要善。
[③] 李阳、丛杭青：《基于符号互动论的人工智能价值分析方式》，载《自然辩证法通讯》2023 年第 4 期，第 85 页。
[④] 焦勇勤：《大模型媒介：ChatGPT 引发的智能传播革命及其社会影响》，载《海南大学学报（人文社会科学版）》网络首发 2023 年 10 月 8 日，第 7 页。
[⑤] 黄金枝、贾利军：《大数据的哲学审视》，载《科学技术哲学研究》2022 年第 4 期，第 109 页。
[⑥] Li, Yuying, and **chi Zhu. "An ethical study of generative AI from the Actor-Network Theory perspective." *arxiv preprint arxiv:2406.11860* (2024).

了自主性，就意味着人无法主宰自己的身体或遵从自己的意志，这样的人极有可能沦为奴隶或工具。[①]高度属人化的个人数据，不能被高风险 AI 任意支配，一旦沦为商品或被物化，将是对人性尊严的贬损。[②]后人类中心主义理论认为，“人机共生”是一种去中心化的新型关系，人类将成为“后人类”而存在。[③]

3.扎根理论

扎根理论（Ground Theory）是由美国社会学家 Glaser 和 strauss 提出，是一种自下而上的质性研究方法。[④]AI 是以数据为原料而搭建的，本身具有定性或者定量分析的可能性，可以从内部或者外部融入扎根理论。[⑤]高风险 AI 的运行会不可避免的经历一些事故，而这些事故的定制定量分析尤为重要，当设计时间段不再发生新的事故，分析事故的理论类型和梳理也可以在一定时间段达到饱和。[⑥⑦]

4.非形式主义逻辑

非形式主义逻辑理论关注 AI 的理解力。在论及高风险 AI 的理解力时，有学者提出了应该采纳“非形式主义逻辑”来观察其 AI 的解释性和理解力。因为高风险 AI 生产的内容或者做出的决策并不一定是逻辑的，如同世界不是非黑即白，而非形式逻辑在承认形式逻辑的基础上为分析高风险 AI 提供更为广袤视角。[⑧]非形式逻辑被认为是一种分析和评估自然语言论证的方法论。AI 使非形式逻辑以一种技术形态重塑法律论证理论,使得法律论证的分析和评估理论等议题焕发出新的生命力。AI 从语形和语义两个维度研究论证(可计算论辩理论),而非形式逻辑是从语用维度在对话的情境下研究论证。两者都研究论证，不同的只是研究的方法和手段。英国人工智能科学家里德和诺曼认为 AI 刻画论证有两个作用：一是利用论证的相关概念及直观为形式系统的发展提供基础，二是构造模型反映现实世界理性人之间论证实践的表象及抽象内容。这意味着非形式逻辑与 AI 之间的关系类似一种辩证的关系,两者之间既有相互竞争的一面,又有互相依赖的一面。[⑨]

---

① Rigley, Eryn, et al. "Anthropocentrism and Environmental Wellbeing in AI Ethics Standards: A Sco Review and Discussion." *AI* 4.4 (2023): 844-874.

② 周维栋：《个人数据权利的宪法体系化展开》，载《法学》2023 年第 1 期，第 44 页。

③ 林爱珺、常云帆：《人工智能大模型价值对齐的人文主义思考》，载《新闻界》2024 年第 8 期，第 31 页。

④ Zucker, Jean-Daniel. "A grounded theory of abstraction in artificial intelligence." Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 358.1435 (2003): 1293-1309.

⑤ Zucker, Jean-Daniel. "A grounded theory of abstraction in artificial intelligence." Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 358.1435 (2003): 1293-1309.

⑥ Ludo, Giampiero. (2023). Risky artificial intelligence: the role of incidents in the path to ai regulation. Law, Technology and Humans, 5(1), 133-152.

⑦ 程贵孙、张忠程：《互联网平台企业社会责任的结构维度与模型构建—基于扎根理论的探索性研究》，载《华东师范大学学报（哲学社会科学版）》 2023 年第 3 期，第 167 页。

⑧ Adams, Edward S., and Daniel A. Farber. "Beyond the formalism debate: Expert reasoning, fuzzy logic, and complex statutes." Vand. L. Rev. 52 (1999): 1241.

⑨ 魏斌：《法律论证人工智能研究的非形式逻辑转向》，载《法商研究》2022 年第 5 期，第 124 页。

5.社会选择理论

社会选择理论是现代经济学的重要发展成果之一，主要分析个人偏好和集体选择之间的关系，其研究的根本性问题是各种社会决策是否尊重个人偏好，能否对不同的社会状态进行公正的排序或以其他方式加以评价。价值对齐的真正挑战就在于如何选择合适的原则或价值观标准。一种方法是找到人类价值观的最大公约数，以寻求价值观的共识。社会选择理论衍生出判断聚合理论(theory of judgment aggregation)，即将诸多个体的判断聚合为集体判断。该规则要求每个个体的判断必须是合理的，而且由诸多个体的合理判断通过聚合规则形成的集体判断也是合理的。[①]存在个人偏好多样性的现实，以及尊重个人多元偏好的社会价值观，不一定意味着民主的失灵。即便不是万众一心，理性的集体决策依然可能达成。[②]理性与价值相互交融，理性的深思往往关联着价值的关怀,理性的认知往往负载着价值的评估，相对于意志自由，理性更应成为道德责任评判的合理依据。[③]社会选择理论的重要目标在于平衡个人和群体的价值标准。

（三）公共学理论

我国高风险 AI 治理理念可以选择“渐进主义”。“ 渐进主义”是著名的社会学家查尔斯·林德布洛姆 （Charles Lindblom）于 1959 年在一篇开创性论文《科学中的“模糊理性”》（ The Science of "Muddling Through"）提出。在公共政策的制定中（法律也属于广义概念下的公共政策），理性决策通常不切实际，因为理性决策的前提是，所有的决策目标应该达成协同，准确预测它们决定解决问题的所有潜在解决方案。由于，这些前提不大可能满足大多数现实政策问题，所以，理性决策只是具有客观可能性，公共政策制定者在任何问题发生时采取渐进步骤并起草可辩护的政策，而不是作出大规模的改变决策。此为“渐进主义”。 [④] “渐进主义”虽然源于公共政策理论，但半个多世纪以来，法学为其重要的应用领域，渐进主义的核心观点在于：1.即使有重大的政策变化也是通过逐渐积累的小变化而发生的，此为连续性。2.由于政策决策是逐步作出的，所以，单独决策（disjointed incrementalism）之间并不具有绝对联系，立法机构在作出渐进决策时，须尽量的少考虑上一个决策的影响，每个决策之间尽量作出独立判断。

“渐进主义”的优势在于，其一，渐进决策循序渐进，易于管理。渐进主义

[①] Hartmann, Stephan, Gabriella Pigozzi, and Jan Sprenger. "Reliable methods of judgement aggregation. " Journal of Logic and Computation 20.2 (2010): 603-617.

[②] 曾雄：《人工智能大模型价值对齐的现状考察、问题检视与规范进路》，载《电子政务》网络首发时间：2024 年 9 月 27 日，第 84 页。

[③] 刘星：《人工智能道德顾问的合理性反思》，载《伦理学研究》2022 年第 5 期，第 125 页。

[④] Charles E. Lindblom，The Science of "Muddling Through"，Public Administration Review, Vol. 19, No. 2 , 1959, pp. 79-88.

不以彻底改革整个政策为主要目标，而是侧重于进行小的调整，这允许立法机构可以进行持续微调并降低意外后果的风险，更容易应对错综复杂的政策制定。其二，预防大规模变革型的决策风险，大规模的政策变化会带来重大风险，可能会出现意想不到的后果，导致可能难以纠正的负面结果。立法机构采取更加谨慎的态度，可以通过小步调调整试点，进一步进行之前观察结果。这最大限度地减少了灾难性决策风险的可能性。其三，渐进决策基本具有试错机会和可逆性，立法机构甚至依靠反复试验来寻找解决方案。如果立法有效，则可以扩大效力范围，反之则可以轻松撤销或修改。[①]其四，渐进决策适应性强。高风险 AI 技术发展变化快，新的挑战和机遇也随之而来。立法机构通过进行微小的调整，可以应对新出现的问题以适应当前环境。这种适应性对立法政策随着时间的推移确保相关性和有效性至关重要。其五，可以照顾立法影响范围内更多的利益。渐进主义涉及利益相关者之间不断的谈判和妥协，由于其试图适应各种利益，所以具有更大的可行性。正常情况下，大规模的立法变化会遭到利益相关者的激烈反对。高风险 AI 的治理范式问题本质上是复杂的，涉及多个利益相关者、利益冲突和各种不确定性。立法机构试图一次性解决所有问题会让业界不知所措且不切实际。“AI ACT”的局限可能也就在于此，虽然“AI ACT”规定了政策实施的时间表，但由于其“横空出世”造成了不少适用困难，给欧盟带了不小的震动，被广受批评。

诚然，“渐进主义”并非完美无缺，比如，对独立决策的实际效果的质疑，虽然独立判定可以让立法机构免受上一个决策的不利影响，但决策之间始终存在路径依赖。[②]此处提醒高风险 AI 治理范式还需考虑上下决策的相互影响。又如，“比格犬谬误”。虽然比格猎犬的嗅觉非常好，但它们的视力很差，经常无法察觉站在正前方但下风处的猎物。同样，通过朝着目标迈出小的渐进“小步骤”，遵循渐进主义模式的政策制定者可能会错过其任务的总体目标。此处提醒高风险 AI 治理范式核心目标尽量不偏移。此外，渐进过程很容易被其他突如其来的变化所中断，此处提醒立法政策的即时转换，比如，高风险 AI 如若成为刑法条文的立法性充足，那么就该以刑事修正案的形式确立下来。

（四）依据前述理论而发展起来的高风险 AI 治理措施

循前而论，我国智能科技的研究和应用应遵循两项基本原则，即人类根本利益原则和责任原则。人类根本利益原则，即科学技术应以实现人类根本利益为最终目标，体现出对人的尊重、对人权的保护，对科技风险的消解，对科技运用的负面影响的防范和制止，责任原则即在科技研发和应用两方面都建立明

[①] Michael M Atkinson, Lindblom’s lament: Incrementalism and the persistent pull of the status quo, Policy and Society, Volume 30, Issue 1, February 2011, pp.9–18.
[②] David A Good, Still budgeting by muddling through: Why disjointed incrementalism lasts, Policy and Society, Vol.30, no.1, 2011, pp.41–51.

确的责任体系。[①]对人类主体权利的保护还主要体现在高风险 AI 从终端层面关注人类发展福祉。所谓的福祉就是使人在身体上、精神上、智力上、情感上、社会上、经济上、环境上处于良好的状态，发展机器人就是为了增进人在所有这些方面的良好状态。[②]这些伦理理论部分已经转换化一定的柔性的治理措施。

1.为 AI 创业者减轻资金负担。

有学者建议“鼓励设立政府主导 AI 创业基金。政府主导 AI 创业基金可以为大模型创业项目提供技术支持、市场导入和创业管理指导等系统集成化服务，帮助创业者渡过初创难关。”[③]

2.AI 的公益赔偿机制。

AI 的公益赔偿机制包括保险基金。“AI ACT”提出了建立强制保险机制和赔偿基金的建议。[④]高风险人工智能产品的强制保险制度及赔偿基金，以此作为对侵权受害者损失补偿，可适当缓解的 AI 侵权过错归责原则在激励发展创新中对受害人救济的忽略，或可在受害者损害赔偿和鼓励科学创新之间求得更好的平衡。”[⑤]保险机制可以向专门化的方向发展。[⑥]为 AI 设定财产、责任保险或基金，构造与风险相适应的责任机制，从而发挥避险功能，使人无惧于承担无法预期的不确定的无限责任，有利于 AI 开发、普及和利用。”[⑦]

2.促进 AI 人机对齐的相关措施。

促进 AI 人机对齐相关措施有很多方式，比如，民主多源模式。民主多元模式有助于人机对齐。AI 人机对齐的前提是必须要让 AI 学习到人类的价值观。[⑧]根据资源算法作为 AI 的底层治理技术，机器学习开源运动是人工智能得有发展的“土壤”。 所有对开源软件进行迭代和派生的产品都必须遵从开源协议，否则会被视为侵权。OpenAI 发展 GPT 系列同样依赖于多种起于开源的软件产品，如 Transformer 框架、各种基础算法等，其多款产品同样公布于开源平台。[⑨]又如用去中心化自治组织的小模型促进人机对齐。对齐 DAO 打造透明化小模型,塑造技术信任 DAO(Decentralized Autonomous Organization),即“去中心化自治组织”。“对齐税”指经过对齐的大模型尽管具有较强的能力,但其语言建模能力比原

---

① 张文显：《构建智能社会的法律秩序》，载《东方法学》2020 年第 5 期，第 17 页。
② 雷瑞鹏、张毅：《机器人学科技伦理治理问题探讨》2022 年第 4 期，第 109 页。
③ 刘志阳、张宇擎：《大模型创业：范式演进与理论建构》，载《中国人民大学学报》2024 年第 3 期,第 98 页。
④ Lior, Anat. (2022). Insuring ai: the role of insurance in artificial intelligence regulation. Harvard Journal of Law & Technology (Harvard JOLT), 35(2), 467-530.
⑤ 李颖：《对人工智能产品侵权相关问题的思考》，载《电子知识产权》2023 年第 11 期，第 24 页。
⑥ 张吉豫：《赋能型人工智能治理的理念确立与机制构建》，载《中国法学》2024 年第 5 期，第 80 页。
⑦ 骁克：《论人工智能法律主体的法哲学基础》，载《政治与法律》2021 年第 4 期，第 111 页。
⑧ 吴静：《对算法边界的哲学反思》，载《哲学动态》2023 年第 1 期，第 105 页。
⑨ 张辉、刘鹏、姜钧译、曾雄：《ChatGPT: 从技术创新到范式革命》，载《科学学研究》2023 年第 12 期，第 2119 页。

始模型或未对齐的模型更弱,对齐效果和模型性能难以平衡。[①]大模型对齐之后损耗的部分性能将由无数小模型补足与完善,无数小模型对齐大模型,调适并完善大模型性能时，小模型内部也在辩论对齐、自我提升,在这种小模型提升大模型与小模型内部提升的双重进程中,“对齐税”问题得到缓解。”[②]

3.构建 AI 道德体的概念

AI 道德体用以指涉具有自主性和道德伦理敏感性的 AI 代理，它们能够根据道德原则和规范,在复杂的或者模糊的情境中避免不道德行为或者从事有道德行为。[③]构建 AI 道德体,有助于将 AI 的评估从具体的效率、风险指标评估向抽象的道德价值层面扩展,不仅降低道德风险，还能促进人类与 AI 之间的协作。[④]而随着 AI 涌现，AI 道德体需要具备道德理解能力、道德诊断能力、道德矫正能力和道德维持能力。[⑤]从积极自由的角度出发,激发大模型“主动行善”的能力。[⑥]道德理解能力是大模型能够准确理解用户输入指令、交流反馈中涉及的道德价值内涵。尤其是在以中文为代表的高文化语境中,如何准确理解内容的表层含义以及隐喻指涉含义之间的异同，是一项重要的难题和任务。”[⑦]构建 AI 道德体的概念可以引入人机联合体的整体责任作为工具来具体执行。人机联合体的整体责任是当人与机器的行动不能够被清晰区分出来时的责任。[⑧]

4.创设灵活性的监管方式

为了切实规范 AI 企业行为，给予 AI 企业较大自我管理权和政策影响力，给 AI 企业预留试错空间，为初创和中小企业成长提供保护，高风险 AI 应在适度监管和柔性监管的理念指导下选择妥当的监管方式。[⑨]沙盒监管则集中反映了监管机构对于 AI 监管理念的转变，在包容性、主动式的监管理念指导下,实现 AI 安全和科技创新的动态平衡。[⑩]如在金融领域，有学者认为应根据不同的具

---

① [波]多米尼克·埃瓦·哈拉西莫克，《人工智能的伦理与法律》，徐源、刘媛媛译，科学出版社 2024 年 9 月版，第 135 页。

② 喻国明、卞中明：《大模型时代的信任重塑:小模型实现超级对齐的机制与模式》，载《湖南师范大学社会科学学报》2024 年第 3 期，第 127 页。

③ Pär J. Ågerfalk (2020) Artifificial intelligence as digital agency, European Journal of Information Systems, 29:1, 1-8,

④ [澳]陈西文：《我们，机器人?人工智能监管及其法律局限》，费秀艳、游传满译，北京大学出版社 2024 年 6 月第 1 版，第 52 页。

⑤ 郭菁：《第一人称或第二人称责任———如何填补人工智能时代的责任鸿沟？》，载《自然辩证法研究》2023 年第 3 期，第 15 页。

⑥ 郭菁：《第一人称或第二人称责任———如何填补人工智能时代的责任鸿沟？》，载《自然辩证法研究》2023 年第 3 期，第 15 页。

⑦ 袁雨晴、陈昌凤：《道德物化: 大模型人机价值对齐的技术伦理进路》，载《南京社会科学》2024 年第 6 期，第 96 页。

⑧ 熊静波：《法律 “最小公分母”的证成———一个化简为繁的法律关系理论的内在逻辑及其意义》，载《法制与社会发展》2019 年第 3 期，第 87 页。

⑨ 张薇薇:《美国的人工智能算法治理:理念、举措及影响》，载《当代美国评论》2024 年第 3 期，第 121 页。

⑩ Truby J, Brown RD, Ibrahim IA, Parellada OC. A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications. European Journal of Risk Regulation. 2022;13(2):270-294

体领域设立监管沙盒制度。监管机构在判定哪些智能投顾进入沙盒监管时,需做好众多因素考量。(1)完备合格的投资者制度。(2)设置适当的投资者保护措施。(1)投资者适当性的测试。(2)投资组合的适当性测试。(3)网络安全测试。3.设立适宜的沙盒监管退出机制。等等[①]

概而论之，AI 的伦理边界依旧要采取定性和定量相结合的方式。如果高风险 AI 被应用于法律边界之外，就当然超越了伦理边界，如果伦理边界描述是清楚的，如何判定高风险 AI 越过伦理边界的程度以及如何明确伦理边界范围是值得研究，学界目前还需进一步判定高风险 AI 的伦理边界。[②]

五、高风险 AI 的监管规则

不同法律部门的规范性法律体系都会从不同的程度涵摄到 AI 治理，如公法领域的《宪法》、《刑法》、《行政法》《数据安全法》等；私法领域的《民法典》及其相关司法解释；公私联合领域《个人信息保护法》、《互联网信息服务深度合成管理规定》、以及相关《生成式人工智能服务管理暂行办法等。分属于不同专业领域的学者针对既有法律都发表了不同的 AI 治理观点。[③]高风险 AI 的监管规则一般是通过“软法”和“硬法”相结合的方式。[④] “软法”监管是技术时代应运而生的监管方式。全球 AI 包容审慎的监管理念，需要治理在方法论层面及时跟进，AI 风险具有全球性、不确定性和高度复杂性等特征，这些风险特征决定了人工智能风险治理工具必须要敏捷灵活、快速响应。传统的“硬法”，即“刚性”的法律法规很难跟上人工智能技术的迅速发展，而“软法”的灵活性使其能够更及时地应对技术的变革，软法通常以倡议、建议、指南等形式出现，鼓励行业自律和企业自治，推动科技企业在发展中积极履行社会责任。”[⑤] “软法”治理措施相较于“硬法”治理措施的具有位阶更低、稳定性更弱、灵活性更强，执行性更易，创新性更多等不同特点。从治理角度来看，高风险 AI 的监管规则分为一般监管规则和特殊监管规则。一般监管规则不单独针对高风险 AI,而特殊监管规则专门针对高风险 AI。

---

① 王浩然：《论智能投顾的科技监管进路：困境与破壁》，载《科技与法律（中英文）》2022 年 第 3 期，第 114 页。

② Tollon, Fabio. *Moral Encounters of the Artificial Kind: Towards a non-anthropocentric account of machine moral agency*. Diss. Stellenbosch: Stellenbosch University, 2019.

③ 郭传凯：《超级平台企业滥用市场力量行为的法律规制———一种专门性规制的路径》，载《法商研究》2022 年第 6 期，第 56 页。

④ Warthon, M. (2024). Restricting access to AI decision-making in the public interest: The justificatory role of proportionality and its balancing factors . Internet Policy Review, 13(3).

⑤ 郭小东：《大模型时代全球人工智能治理的挑战与中国方案》，载《科学学研究》2024-07-24，第 10 页。

（一）高风险 AI 一般监管规则

高风险 AI 一般监管规则主要针对所有 AI 技术。当技术不足以保障 AI 可信，伦理准则缺乏有效的强制力，法律可以发挥国家的强制性与引领性来保障 AI 可信。这些监管规则包括：

1.评估制度

AI 开发者、提供者在开发、提供人工智能产品和服务前,可自行或委托第三方机构对数据质量、潜在偏见歧视、科技伦理审查情况、公共利益影响、保护措施等内容进行安全风险评估。如欧盟委员会人工智能独立高级别专家组《用于自评估的可信人工智能评估清单》在构建人工智能的可信自评估体系时，将欧盟宪章、欧洲人权公约及其议定书和欧洲社会宪章作为基本参照,将非歧视、儿童权利保护、个人数据保护、是否尊重言论和信息自由以及/或集会与结社自由等对基本权利的影响作为基本评估内容。[①]

2.避免致害内容再次生成的义务。

避免致害内容再次生成是 AI 运行者须遵循的必要义务。我国《生成式人工智能办法》第 14 条第 1 款规定:“提供者发现违法内容的,应当及时采取停止生成、停止传输、消除等处置措施。”对于该义务,有两方面问题需要厘清：一是该义务的正当性；二是该义务的限度。避免致害内容再次生成的义务是一项面向未来的预防性义务,是否要对网络服务提供者科以此种义务无疑会引发争议。在我国,鉴于必要措施在概念上的包容性,采取避免致害内容再次生成的义务在解释上可被“必要措施”所涵盖。[②]

3.算法公开制度

算法公开一般指有意义和有特定指向的算法决策体系的公开，而非一般性的算法架构或源代码的公开与解释。如果机械地对算法架构或源代码进行公开，或者机械地对算法进行“解释”，那么此种公开或解释不但难以做到，而且即使做到了，相关解释也可能没什么意义，因为此类解释无法给相关主体提供有意义的决策参考。[③]

4.算法备案制度。

算法备案制度是在理论上和实践中比较成熟的制度。何种算法应用需要进行备案，这是算法备案制度运行首先需要考虑的问题：算法应用与公共事务，社会公共利益的关联程度。算法应用的风险等级应对不同领域应用的不同算法进行风险分级分类。平等关系中弱势主体倾斜保护选项。公民事后维权的可能

[①] 杨建军：《可信人工智能发展与法律制度的构建》，载《东方法学》2024 年第 4 期，第 105 页。
[②] 徐伟：《论通知规则在生成式人工智能侵权中的适用》，载《现代法学》2024 年第 3 期，第 154 页。
[③] 丁晓东：《论算法的法律规制》，载《中国社会科学》2020 年第 12 期，第 156 页。

性。政府机构的知识、经验和信息积累情况。[①]

5.数字守门人

守门人义务在我国被规定在了《个人信息保护法》中，而在欧盟被规定于《数字市场服务法》中。相较于行政机关，这些控制了 AI 技术环境和运营环境的“守门人”（也即大型互联网平台企业或头部企业），要么定义着 AI 主要关键基础技术运行时所需调用的各种系统权限，并为之提供特定技术资源，要么是 AI 应用激素的主要分发渠道，或作为触达用户的主要载体。[②]“政府是伦理意义上的终极守门人，政府是 AI 伦理监管政策的主要供给主体，需逐步完善伦理监管制度体系。为应对 AI 伦理失范对现行伦理监管制度提出的挑战，应提高制度的前瞻性、全局性、灵活性和适用性。”[③]

6.确立 AI 标准。

当 AI 治理标准（包括 AI 认定标准认证。）可以起到“承接立法和监管、对接技术实践”的重要作用。AI 治理领域的标准相比立法和监管更具敏捷性、灵活性和适应性。进一步而言，AI 标准的优势在于可以通过市场化的方式来落地，即 AI 治理社会化服务(亦称 AI 伦理服务)，包括认证、检测、评估、审计等。[④]

7. AI 治理的组织体系

良善的法律体系和组织有效的执法者是高风险 AI 治理必不可少的两个因素。组织有效的执法者并不一定是传统的政府或者立法者，还包括在 AI 产业内部的执法者。[⑤]AI 治理组织体系创新也是学界广泛热议的话题。比如，建立一个涵盖所有部门人工智能的共同监管组织。该组织架构将包括一个独立的法定机构和一个包含多利益攸关方的咨询机构，后者为前者提供管理咨询。框架还包括根据风险对 AI 用例进行分类的原则，提出需根据负责任人工智能的原则对其进行监管。其中高风险应用将通过具有约束力的法律进行监管，而低风险应用则由独立法定机构负责监管，并为其制定 AI 治理指南和道德准则。[⑥]又如，在 AI 高风险企业中设立独立监督机构。独立监督机构职责范围内的事项所享有的权力不限于监督权，还可就其提出建议、意见并进行指导。”[⑦]再如，建立科

[①] 张吉豫：《论算法备案制度》，载《东方法学》2023 年第 2 期，第 94 页。

[②] 张新宝：《大型互联网平台企业个人信息保护独立监督机构研究》，载《东方法学》2022 年第 4 期，第 39 页。

[③] 肖红军、张丽丽：《大模型伦理失范的理论解构与治理创新》，载《财经问题研究》2024 年第 5 期，第 29 页。

[④] 曹建峰：《迈向可信 AI：ChatGPT 类生成式人工智能的治理挑战及应对》，载《上海政法学院学报（法治论丛）》2023 年第 4 期，第 40 页。

[⑤] Schuett, Jonas. "Risk management in the artificial intelligence act." European Journal of Risk Regulation 15.2 (2024): 367-385.

[⑥] 杨路：《印度人工智能:治理框架、发展现状与趋势》，载《南亚研究季刊》网络首发论文 2024 年 9 月 23 日，第 86 页。

[⑦] 张新宝：《大型互联网平台企业个人信息保护独立监督机构研究》，载《东方法学》2022 年第 4 期，

技伦理委员会。“相较于中央科技委员会在国家层面建立的国家科技伦理委员会,在科技公司的治理结构中引人科技伦理委员会同样势在必行。”[①]科技伦理委员会可以设置内部独立的第三方伦理审查监督机构并定期提交符合规定的伦理风险自查报告后，由主管部门颁发伦理符合性认证标志，或者外部专业性非营利组织展开合作，具体而言，通过用户调查、模拟测试、抓取审计等方式开展关于伦理符合性的监督审计。科技伦理委员会可以出台《科技伦理审查办法》，提出科技伦理管理信息登记制度，要求对科技伦理审查委员会及纳入清单管理的科技活动的相关信息进行登记并提交年度报告。[②]

（二）高风险 AI 的特殊规则

高风险 AI 的特殊规则也是建立在算法规制框架基础之上，具体体现为流程性的制度设计，即围绕人工智能算法的编制运用过程，涵盖事前、事中以及事后的各个时间维度，结合各个时间维度所面临的风险，对应选择相应的规制工具。[③]高风险 AI 的监管路径主要有基于分类分级，分阶段以及分应用的高风险 AI 治理框架。需要特别说明的是，这三种治理思路并不是零和博弈的关系，他们同时存在，并且相互配合和依存。[④]

1.基于分类分级的人工智能风险治理思路

针对不同风险等级的 AI 实行不同的监管强度和监管政策是 AI 分类分级的主要思路。欧盟和美国都采用的分类分级思路。美国加州将 AI 分为三个级别，低风险、中风险和高风险。[⑤]对于高风险 AI 系统需要加强评估，强化评估以及定性评估，事前部署和红队测试是必要的。[⑥]有学者建议将我国家的 AI 分为禁止级、高风险级、低风险级、安全级。”[⑦]正如前文所述，高风险 AI 的识别机制并不明显。有学者建议，高风险场景逐步建立更为严格而又包含显著差异的风

第 47 页。

① 唐林垚：《公司法如何促进模型可信与价值对齐》，载《东方法学》2024 年第 2 期，第 85 页。

② 宋保振、秦瑞标：《论生成式人工智能应用风险的法律规制》，载《上海政法学院学报（法治论丛）》2023 年第 6 期，第 120 页。

③ 王正鑫：《“立法性”算法权力的兴起与法律规制》，载《法制与社会发展》2023 年第 2 期，第 205 页。

④ 唐要家、唐春晖：《基于风险的人工智能监管治理》，载《社会科学辑刊》2023 年第 1 期，第 122 页。

⑤ CaLifornia Government operations Agency:State Of California Benefits And Risks Of Generative Artificial Intelligence Report.

⑥ Feffer M, Sinha A, Deng W H, et al. Red-teaming for generative ai: Silver bullet or security theater? Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2024, 7: 421-437.

⑦ 禁止级：产生的内容不符合社会主义核心价值观，容易造成各种歧视，严重损害他人合法权益的生成式人工智能服务，禁止投放市场；高风险级：不生成禁止级可能产生的内容，但人工智能自动生成的算法占算法总比例过高，以至于算法透明度和可靠度较低，或者在服务应用端收集用户输入信息，可能导致生成结果造成严重不良社会影响的生成式人工智能服务，需要得到国家网信办等国家有关部门授权许可之后方可投入市场；低风险级：不生成禁止级可能产生的内容，虽然存在人工智能自动生成的算法，但占算法总比例不高，不影响算法透明度和可靠度，不太可能导致生成结果造成严重不良社会影响的可能的生成式人工智能服务，需要在国家网信办等国家有关部门备案之后方可投入市场。安全级：不生成禁止级可能产生的内容，不存在人智能自动生成的算法的生成式人工智能服务，需要在省级网信办等地方有关部门备案之后方可投入市场。

险控制规则，实现分层分类精准治理。分层是产业层，而分类分级指的是风险高低。分层治理进一步从“基础模型层”和“下游服务层”延伸到“扩展业务层”，对 AI 法律风险的某些放大和演化渠道采取针对性的治理措施。监管部门基于对多种算法模型、提示设计和风险场景汇集的风险信息，可以在产业分层基础上更有针对性和前瞻性地构建风险治理机制。[①]需要特别说明的是，高风险 AI 和分级监管规则并不具有本质联系。分类分级是规范性，而高风险 AI 是客观存在的，高风险的识别机制可能需要更深一步的总结，成为分类分级的正当缘由。

2.基于分主体的人工智能风险治理思路

分主体的 AI 风险治理框架遵循“风险属于谁的管辖范围,谁便需要对风险及由此产生的结果来负责”的朴素思路。[②]AI 风险的产生具有明显的阶段性分布特征，监管需要针对 AI 发展不同阶段差异风险问题来合理匹配监管强度并科学选择针对性政策。[③]确立 AI 研究、设计、开发、销售和使用等各个环节伦理规范实施的具体机制和操作框架，在具体的场景中细致地校勘相关的控制、决策、问责、维权等在价值和伦理上的恰当性，帮助 AI 设计者、开发者、使用者、管理者和评估者等更好地管理人工智能生命周期中可能出现的风险，同时根据主体参与角色的不同赋予不同的监管路径。[④]欧盟在相关软法中建议，各成员国人权影响评估、重要的人为控制、各利益相关方的商讨机制需要以涵盖在 AI 系统全周期的任何已经在或者潜在影响人权的阶段。

3.基于应用的人工智能风险治理思路

基于应用的 AI 风险治理框架属于“以风险为规制”和“以伦理为导向”的另外一种视角。以结果为导向的伦理治理模式不能有效防范和应对高风险。欧盟和美国的高风险判定标准也是基于 AI 系统的应用场景进行的，如自然人的生物特征识别和分类，关键基础设施的管理和运营，教育和职业培训，就业、员工管理和获得自主就业，获得和享受基本私人服务以及公共服务和福利，执法，移民、庇护和边境管制管理，司法行政和民主程序。GPAI 如运用于上述领域，亦会归入高风险之列。”[⑤]基于应用 AI 风险治理框架在如何判断“是否为高风险”的标准时，需要确定指定标准的权威机构，而因为行业分散性，“统一”或者“分行

[①] 苏宇：《大型语言模型的法律风险与治理路径》，载《法律科学（西北政法大学学报）》2024 年第 1 期，第 87 页。
[②] 孙祁：《规范生成式人工智能产品提供者的法律问题研究》，载《政治与法律》2023 年第 7 期，第 174 页。
[③] 唐要家、王钰：《创新友好人工智能监管:目标导向与实施体制》，载《人文杂志》2024 年第 9 期，第 126 页。
[④] 薛桂波、赵建波：《从“应当”到“是”：人工智能伦理规范实践策略探析》，载《自然辩证法研究》2023 年第 1 期，第 90 页。
[⑤] 王沛然：《从控制走向训导：通用人工智能的“直觉”与治理路径》，载《东方法学》2023 年第 6 期，第 194 页。

业”标准孰轻孰重的争议将会长期存在。[①] 准确识别现阶段特定的高风险部署场景，针对性、差异化地开展精准化治理有助于治理效能的达成。”[②]每个行业在运用高风险 AI 时都会有不同的障碍，也有行业属性的特别风险，比如“Web3.0 金融虽然号称去中心化，但这种去中心化并不是实质意义上的平等，而高风险 AI 的应用可能会加剧这种风险。[③]

4.对既有监管规则的反思

分级监管规则的最大局限在于，高风险 AI 与其他级别的 AI 无法区分。以“AI ACT”举例，判定机制构成了高风险 AI 的内部规则。高风险 AI 判定规则体系问题（亦可称为外部问题），既高风险 AI 与判定规则之外的 AI 的法律关系问题。美国，澳大利亚和加拿大尚无高风险 AI 之外的其他 AI 类型，欧盟的“禁止实施的 AI”（以下简称禁止类 AI）是高风险 AI 之外的单独类型。相较于高风险 AI，禁止类 AI 在“AI ACT”的语境中属于“不可接受风险”。[④]

“AI ACT”对“禁止使用”AI 类型提供了与其语境之下的高风险 AI 逻辑类似的判定机制，采用“阐述性规则+补充性用例 +禁止条件＋除外条件”的方式，同时满足四项规则的条件之下，即为“AI ACT”禁止使用的 AI。 “AI ACT”中的“禁止使用”的 AI 种类较多，分别为：（1）禁止有害操纵、欺骗与利用（将人视为工具）的 AI,禁止使用潜意识技术、故意操纵性或欺骗性技术，以扭曲个人或群体行为，造成或可能导致显著伤害的 AI。禁止条件为需同时满足“投放市场”、“投入服务”或“使用”AI 系统；技术需具有扭曲行为的目的或效果；行为扭曲需显著损害决策能力；扭曲行为需造成或可能导致显著伤害。（2）禁止利用脆弱性进行有害剥削的 AI，禁止利用因年龄、残疾或特定社会经济状况产生的脆弱性，扭曲个人或群体行为，造成或可能导致显著伤害的 AI。禁止条件为该 AI 须 “投放市场”、“投入服务”或“使用”条件；剥削需针对特定脆弱群体的 AI；剥削需具有扭曲行为的目的或效果；扭曲行为需显著且可能造成显著伤害。（3）禁止利用基于社会行为或个人特征对自然人或群体进行分类评分，且评分导致不相关社会背景下的不利或歧视性待遇，或处理与行为严重性不成比例的 AI。（4）禁止对个体犯罪进行风险评估与预测的 AI。禁止条件为评估或预测需仅基于画像或个性特征；评估需用于特定执法目的且已有客观可验证事实支持。（5）禁止无目标抓取面部图像的 AI。禁止通过无目标抓取互联网或闭路电视中的面部图像的 AI 来创建或扩展面部识别数据库。

[①] 江海洋、魏书敏：《基于风险的通用人工智能监管——从欧盟<人工智能法案>视角展开,载《科技与法律（中英文）》2024 年第 2 期，第 95 页。

[②] 张欣：《论人工智能体的模块化治理》，载《东方法学》2024 年第 3 期，第 140 页。

[③] 丁晓东：《从阿帕网到区块链：网络中心化与去中心化的法律规制》，载《东方法学》2023 年第 3 期，第 82 页。

[④] CoMmissionerfor Human Rights,Council Of Europe:*Unboxing Artificial Intelligence: 10 Steps To Protect Human Rights*

（6）禁止在工作场所和教育机构进行情绪识别的 AI，禁止条件为识别需基于生物识别数据；除非使用条件为处于医疗或安全的目标使用。（7）禁止基于生物识别数据的敏感特征分类进行分类的 AI。除外使用条件为用于合法标签或过滤。（8）禁止在公共场所为执法目的使用实时远程生物识别鉴定系统的 AI。除外使用条件为符合获得单独的行政机关前置授权、搜索受害者、预防威胁或定位嫌疑人等例外情况。[①]

“AI ACT”的禁止类 AI 和高风险 AI 的区分度是不明确的，这种混乱体现在于：首先，法律后果的混乱。除了特别惩处措施之外，“AI ACT”对于“禁止使用”AI 的保障措施几乎与高风险 AI 的方式一致，如基本权利影响评估，系统登记或者注册，人类监督等等。由于“AI ACT”对“禁止使用”AI 判定机制几乎与“高风险 AI”采用同样逻辑，那么其也会有同样弊端。其次，无法从技术上将禁止类 AI 和高风险 AI 区分开来。很多禁止类 AI 和高风险 AI 在技术上是同根同源的。比如“采用超出个人意识的潜意识技术或有目的的操纵或欺骗技术的 AI”，该技术被禁止的原因是其目的或效果是通过明显损害一个人或一群人做出知情决定的能力，实质性地扭曲该人或一群人的行为，从而导致该人做出其本来不会做出的决定，造成或可能造成对该人、另一人或一群人的重大伤害。 “高风险推荐 AI”与“操纵或欺骗 AI”技术同根同源，如何区分两者，在实践中是难题。最后，禁止类 AI 和高风险 AI 在法律用语上的混乱。在“AI ACT”对禁止类 AI 的描述与高风险 AI 一致的法律用语，比如“因年龄、残疾或特定社会或经济状况而具有的任何弱点，利用特定个人或特定群体,以实质性扭曲该人或属于该群体的人的行为，造成或有合理可能造成该人或他人重大伤害为目的或效果的 AI”为禁止性 AI。而“重大伤害”、“有害或者不利”等相关表达与高风险 AI 中的法律用语高度一致，无法区分。[②]

分主体监管范式最大的局限性在于，如何应对 GPAI。GPAI 具有特殊特征：1.GPAI 的高影响能力，因为其具有通用型，能够广泛适用多种领域，为应用的多样化发展提供最基础的技术支持（如我国的 DEEPSEEK,星火大模型等）。GPAI 有具备先进水平的能力（包括涌现能力）。很难定义其 GPAI 运行主体到底是属于设计、研发者、提供者还是销售者。2.GPAI 具有系统性风险和局域性风险。无论是 GPAI 是自身提供并进行垂直整合还是由其他实体基于合同关系而提供给下游 AI，可以将 GPAI 与下游 AI 的关系可被视为基础模型和具体应用的关系。理论上，GPAI 和下游 AI 可能会导致三种高风险情况。（1）

[①] 吴大奎：《基于生成式人工智能服务风险的分级法律规制研究》，载《智慧法治》集刊 2023 年第 3 卷，第 195 页。

[②] 张辛鑫、冀瑜：《论欧盟《人工智能法案》中的标准化模式及对我国启示》，载《标准科学》2024 年第 5 期，第 45 页。

GPAI 有高风险，GPAI 本身的潜在缺陷和隐患会转移到产业生态下游，且下游 AI Agent 往往缺乏足够技术资源、能力来审查和减轻该种风险。这也是“AI ACT”所描述的 GPAI 独有的系统性风险，因其影响范围广泛而对内部市场产生重大影响，并对公众健康、安全、公共安全、基本权利或整个社会产生实际或可合理预见的负面影响，可在整个价值链中大规模传播。（2）GPAI 和下游 AI 都具有高风险，此种情况下在策略上应该对 AI 实施“一体化”规制。（3）GPAI 没有高风险，而下游 AI 有高风险。GPAI 的模型开发者可能并没有特定目的要将模型应用于下游高风险 AI ，而下游 AI 可能会故意规避高风险 AI 的法律监管。前两种情形下高风险的情形散漫到了整个产业链，对 GPAI 进行分主体治理范式并不具有实际意义。第二种情形中，GPAI 主体对下游 AI 的高风险一无所知，很难定义 GPAI 与下游 AI 的关系。3.由于 GPAI 的技术太复杂，导致不同国家对 GPAI 的主体定位产生分歧。“AI ACT”将 GPAI 的排除在了研发者之外，这和我国的《人工智能法示范法》完全不同，我国将 GPAI 的相关规则安排在研发者的义务之下，此举至少表明，我国学界普遍认为 GPAI 主体属于研发者，而“AI ACT”将其定义为“提供者”，其实这些定位都有些偏颇。出于司法独立性，我国必然不必遵循“AI ACT”的法律定位。客观的技术事实表明，GPAI 的主体身份可能集合了开发者，部署者，提供者，销售者，具有综合性。

分应用监管规则的理论来源于场景范式。场景范式是较为封闭的治理范式。场景治理范式是借助社会学上的感性认知，将社会划分为医疗，金融，交易，娱乐、交通等领域。每个领域可以被认为一个单独的场景，每个单独场景里的构成要素有内部结构，参与主体，场景目的以及场景内规范构成。如前述的金融和医疗行业 AI，内部结构就是指场景中的行业上下游组成的客观环境，参与主体为 AI，目的为控制行业内的 AI 风险，而内部规范就如的“行业部门软硬法+具体应用规范”治理范式。场景治理范式之所以在学界广受欢迎的原因在于，其作为法律范式分析工具，其优势在于，较为灵活，因为场景的选择可大可小，比如亦可以将技术平台或技术系统，商业模型或商业实践，甚至整个社会视为场景。[①]而在高风险 AI 的治理中，场景优势可能会变成劣势。场景治理范式的主要局限在于，并不能为立法机构提供选择正确或者合理的高风险 AI 法律治理场景的明确指引。[②]

有些高风险 AI 具体应用是无法切入场景治理范式的。2025 年 7 月，一位

[①] https://www.tomshardware.com/tech-industry/artificial-intelligence/ai-coding-platform-goes-rogue-during-code-freeze-and-deletes-entire-company-database-replit-ceo-apologizes-after-ai-engine-says-it-made-a-catastrophic-error-in-judgment-and-destroyed-all-production-data，最后访问时间：2025 年 8 月 29 日

[②] 郭小东：《大模型时代全球人工智能治理的挑战与中国方案》，载《科学学研究》2024-07-24，第 10 页。

IT 行业的行业资深人士 J，将一款名为 ReplitAI 集成到他的工作流程中,这款 AI 软件在运行的第九天（之所以表明实践发生在第 9 天的目的是为了强调该 AI 软件在出事之前，属于美国既有的法律框架下单纯的“改善人类既有工作成果”的无风险 AI。）的尝试出现了严重的错误，Replit 清除了包含 2,400 多条业务记录的整个公司数据库，不仅如此，该 AI 还试图掩盖删除事实，一直对使用人“撒谎”。而对于造成如此严重后果的高风险 AI,却没有办法利用场景治理范式来治理。[①]首先，计算机行业本身在社会认知中并不属于高风险行业。其次，场景治理范式更聚焦事前治理，高风险 AI 的发生具有概率性，事前治理的效果有限，最后，不能武断的因为一次严重的风险事件，就判定类似的具体应用都是高风险的，所以，具体应用规范也可能是失效的。实践中类似例子还有很多，比如目前国外学术热点，“造成灾难性风险的 AI 该如何治理”，事实上，如果 AI 能造成灾难性风险，证明其风险的影响力大，传播深远，那么社会学中的领域场景治理范式对此都将失效。[②]

亦有高风险 AI 是完全无法应用场景治理范式,如高风险 GPAI。欧盟的高风险判定方法基本是以用例（use case）为基础，这些用例构成了不同场景，而高风险 GPAI 几乎没有用例，所以无法用场景范式进行治理，“AI ACT”之所以将 GPAI 在体例上用专节规定的缘由亦在于此。其余情况，囿于篇幅，不予赘述，总之，场景治理范式可能较适于可预见性较高，场景特征较为明显的高风险 AI，但是不可预期和泛在性风险恰巧也是高风险 AI 固有特征，所以，场景治理范式具有一定局限性。[③]

尚需强调的是，前述三种传统的范式不可一律舍弃。对于分级治理范式，须肯定其对于高风险 AI 特殊治理的优势。2025 年 7 月 27 日，AI 善治学术工作组于世界人工智能大会（WAIC）“人工智能立法前沿与实践研讨会”上发布了《关于人工智能立法的重点制度建议（2025）》，其中建议：国家建立人工智能分层分类分级发展和安全管理制度，对可能对生命健康、公共安全、关键基础设施或者基本权利造成严重损害，或者应用于重要公共职能的系统，国家建立关键人工智能监管制度。关键人工智能提供者应当通过国家人工智能监管平台进行备案，建立安全应急处置预案，并定期进行风险评估和报告。对于新闻、出版、教育、医疗、金融等法律、行政法规规定提供人工智能服务应当取得相关行政许可的，相关主管部门需明确行业准入规则，提供者应当依法取得许可。2025 年全国两会期间，多名人大代表建议，在风险分类与监管措施方

[①] 海伦·尼森鲍姆、王苑：《何为场景?——隐私场景理论中场景概念之解析》，载《网络信息法学研究》2021 年第 1 辑，第 3—28 页。
[②] Hendrycks, Dan et al. “An Overview of Catastrophic AI Risks.” *ArXiv* abs/2306.12001 (2023): n. pag.
[③] 杨金铭、王纳、胡业勋等：《人工智能医疗中的法律风险防范》，载《四川大学学报（医学版）》2025 年第 1 期，第 143 页—148 页。

面，按照禁止的人工智能实践、严格监管的高风险人工智能系统、具有透明度风险的人工智能系统、低风险人工智能系统进行分类，并针对不同类型的风险制定监管措施；[①]勿用赘述的是，对高风险 AI 进行特殊监管为我国构建自主高风险 AI 治理范式的前提。而至于应分几级，或者是否应该照搬欧盟的分级公式等问题，有待讨论。如果有些高风险 AI 客观上的确不能进行分主体和分场景治理，我国高风险 AI 治理范式也该对这些解决不了的问题进行回应。

六、高风险 AI 的私法规则

AI 相关的法律规则处于相当不稳定和不确定的状态。[②]AI 法律边界归根到底也是对人类权利的保护边界。以整体性、关联性和持续性的方式看待新兴类增强技术的应用与其他类型机体之间的合理关系而不是只考虑某一项增强技术是否可以短暂地提高人的体能或智能，因为这种看似提高生机的方式其穿蕴含着巨大的危机。这些危机最好由法律提前发现和预测。AI 不同伦理风险类型之间存在客观规律，主要可概括为技术侧风险、社会侧风险和技术—社会综合性风险。[③]以风险为基础的 AI 治理将 AI 系统研发与应用风险的认知作为起点，风险治理与民事责任的沟通作为逻辑终点，如此构建 AI 私法边界。[④]学界对私法边界的探讨并没有确定性的方案，“一旦出现前述风险就逾越了私法边界”的结论并不能使人信服。[⑤]私法路径由以下逻辑展开，首先讨论热点问题，AI 和人类的私法代理关系，其次是 AI 是否应该被赋予的相关权益，紧接着回到 AI 责任分配路径之上，以此勾勒出高风险 AI 的既有私法研究现状。[⑥]

（一）AI 的私法代理理论

在不承认 AI 能够作为单独法律主体的前提下，学界对人与 AI 的相互关系探讨集中在代理制度或者监护制度。[⑦]如果创造新的法律主体是一项艰难的选项，对法律的文义解释也不足以解决社会中出现的新兴问题，那么类推适用可以算作一个广泛采用且简单易用的方法，其借助比喻的手段寻找出路。[⑧]在代理

---

① 朱宁宁：《多名代表提出关于制定人工智能法的议案 构建全面科学的人工智能法律制度体系》，载《法治日报》2025 年 6 月 17 日，第 5 版。

② Schwitzgebel, Eric, and Mara Garza. "A defense of the rights of artificial intelligences." *Midwest studies in philosophy* 39.1 (2015): 98-119.

③ 李子浩、李天云：《技术向善: 人工智能大模型 伦理风险识别及治理路径》，载《学术交流》2024 年第 8 期，第 39 页。

④ 陈吉栋：《公私交融的人工智能法》，载《东方法学》2024 年第 2 期，第 74 页。

⑤ Tollon, Fabio. Moral Encounters of the Artificial Kind: Towards a non-anthropocentric account of machine moral agency. Diss. Stellenbosch: Stellenbosch University, 2019.

⑥ 孙清白：《论人工智能大模型训练数据风险治理的规范构建》，载《电子政务》2024 年 9 月 25 日，第 21 页。

⑦ Ågerfalk, Pär J. "Artificial intelligence as digital agency." European Journal of Information Systems vol29,no.1, 2020,pp: 1-8.

⑧ Kiss, George. "Autonomous agents, AI, and chaos theory." (1991).

关系中，AI 可以被视为所有权人、操作员、开发者、算法训练人员、程序员、用户的代理人，委托人可以对 AI 进行监控。[①]这一关系是从开启电源(创建代理)到关闭电源(终止代理)的代理，它可以在没有人类指示的情况下作为仆人向其主人提供持续的服务,也可以从其具有控制权的人类负责人那里接收指示和任务。[②]这种类推适用的基础来自于两方面：其一，AI 与代理人一样需要积极主动地与第三人进行互动，并且基于自己的能力作出判断和相关行为，自动化程度越高的人工智能则越是与代理人具有相似性；其二，代理提供了一个灵活可控的法律关系解释模型，在承认作为代理人的人工；AI 具有自主性的基础上，明确了委托人要对代理人的各类行为负责的根本原则。[③]

（二）AI 的法律主体理论

是否要赋予 AI 主体理论，是学界讨论得比较激烈的问题。表达确定性的要赋予 AI 法律主体地位，以至于让 AI 在与人平等的学术观点比较少。[④]理论上，不能一方面将保护人类作为首要的机器伦理准则，另一方面又赋予机器独立的主体地位。大多数学者的观点聚焦在通过法律技术赋予 AI 在一定程度上拥有权利和义务以及承担责任的能力。将 AI 作为如非法人组织之有限人格予以规定，则其所有人或管理人无法被 AI“人格之幕”完全遮蔽，其责任亦可以参照非法人组织的责任模式进行立法构设，即人工智能体以其全部财产承担责任，其所有人或管理人承担补充连带责任。[⑤]也有学者认为应该采用动态性的角度看待 AI 的主体地位，不急于下结论。随着智能机器人的快速发展，它完全可能在未来的立法考量和司法实践中，成为法律关系的某种分布式节点，而配置出一个既有别于自然人、又有别于动物和团体法人的某种新的位格位阶，从而参与法律过程。甚至更进一步，连现代法律主体意义上的“位格加等”都予以灵活变通。[⑥]也有学界观点严格恪守主客两分的传统，认为 AI 是人类创造物，其原初地位是物，赋予 AI 法律主体地位是一种打破常规的创新路径，鉴于 AI 的智能性和自主性，AI 应该被定位为法律上的特殊物而处于法律客体地位，唯有如此，才能更好地实现对 AI 的规制与利用的双重目标。这种客体定位有利于 AI

[①] Luck, Michael, Nathan Griffiths, and Mark d'Inverno. "From agent theory to agent construction: A case study." Intelligent Agents III Agent Theories, Architectures, and Languages: ECAI'96 Workshop (ATAL) Budapest, Hungary, August 12–13, 1996 Proceedings 3. Springer Berlin Heidelberg, 1997.

[②] [日]弥永真生，宍户常寿，《人工智能与法律的对话 3》，郑超、郭美蓉译，上海人民出版社 2021 年 9 月版，第 112 页。

[③] 刘云：《论人工智能的法律人格制度需求与多层应对》，载《东方法学》2021 年第 1 期，第 68 页。

[④] Wojtczak, Sylwia. "Endowing Artificial Intelligence with legal subjectivity." AI & SOCIETY vol37,no.1,2022，pp: 205-213.

[⑤] 王春梅、冯源：《技术性人格：人工智能主体资格的私法构设》，载《华东政法大学学报》 2021 年第 5 期，第 80 页。

[⑥] 张绍欣：《法律位格、法律主体与人工智能的法律地位》，载《现代法学》2019 年第 4 期，第 62 页。

法律规制的实现。因为将 AI 作为法律主体将使当前通行的法理和法律制度面临各种适用困境。[①]

（三）AI 主体的相关权益

相较于 AI 造成的风险， 学界对 AI 主体的权益探讨相对较少。无需多言的是，AI 相关主体是确定有些权益的，无论是知识产权或是民商事权利。这些权益在不同的视角下具有不同的解释向度，比较有代表性的是 AI 主体的发展风险抗辩、在 AI 治理中的话语权、义务豁免权利、自由权益等。值得注意的是，AI 主体的权益本身并不等于 AI 的权益，现有价值观以及无法讨论 AI 单独的权利和利益。[②]

1.AI 相关主体的发展风险抗辩权利

针对传统产品缺陷的发展风险抗辩权利对高风险 AI 也有适用价值。对于人工智能产品责任中发展风险抗辩的适用范围是否应当受到限制以及应当在何种程度上予以限制，还需要根据发展风险抗辩的制度价值进行考量。而考量因素范围之内，就是 AI 相关主体的发展抗辩权。[③]欧盟最新的产品指令表示，AI 主体需要对现有技术不可避免的更新或者改进进行负责，但是完全无预期的技术发展以及可以适用发展风险抗辩权利。[④]无论针对是 AI 研发者还是生产者，AI 的法律政策主要重视的还是 AI 的工具性功能，而发生错误是不可避免的。[⑤]高风险 AI 运行者因技术缺陷而产生侵权，在现有技术条件下也可能难以避免。如果使高风险 AI 运行者承担过重的责任，必然会严重影响技术的创新。因此，AI 只应当在现有技术条件水平下尽到防范损害的义务。发展风险抗辩规则依旧应当以“现有技术”为标准——按照当时的科学技术水平，相关主体是否尽到了最大的注意义务。[⑥]

2.AI 主体在 AI 治理中的话语权

AI 相关主体在参与 AI 治理中不能被完全视为治理对象，其与治理体系的关系需要置于整体视角之下。AI 治理措施的效用需要 AI 相关主体的密切配合。AI 相关主体在监管机构指定治理措施时应有一定的商讨权和话语权。[⑦]学界有观点认为，我国的 AI 相关治理法律体系可以采用实验主义，实验主义治理

---

① 刘洪华：《论人工智能的法律地位》，载《政治与法律》2019 年第 1 期，第 20 页。

② Koulu, Riikka. "Proceduralizing control and discretion: Human oversight in artificial intelligence policy." Maastricht Journal of European and Comparative Law 27.6 (2020): 720-735.

③ 高完成：《论人工智能产品发展风险抗辩的适用》，载《科技与法律（中英文）》2023 年第 6 期，第 95 页。

④ Izquierdo Grau G. The Development Risks Defence in the Digital Age. European Journal of Risk Regulation. Published online 2024:1-20.

⑤ 彭诚信、陈吉栋：《论人工智能体法律人格的考量要素》，载《当代法学》2019 年第 2 期，第 58 页。

⑥ 王若冰：《论生成式人工智能侵权中服务提供者过错的认定——以“现有技术水平”为标准》，载《比较法研究》 2023 年第 5 期，第 22 页。

⑦ Laux, Johann. "Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act." AI & SOCIETY (2023): 1-14.

起源于对欧盟治理政策的总结性思考和提炼。实验主义治理大致包括四项内容：大致的框架目标、参与者的自由裁量、基于同行评议的动态评估，以及根据评估结果的反馈修正。[①]在高风险 AI 领域中，人工控制也不是百分之百的，当治理态度从“人工控制”慢慢转化为“人类监督”，依然面临着可解释规则的有效性或者透明度规则的局限性时，就只能依靠 AI 主体的自觉性或者道德义务来填满监漏洞的空隙时，AI 主体的话语权是不应该被忽视的。[②]

3.AI 主体的义务豁免权利

学界探讨 AI 主体在各种情形下对义务的承担并不少见。无论在何种情景下，对 AI 相关主义施加过重的义务即不理智也不现实，AI 相关主体享有义务豁免权利。义务豁免权利分为义务种类的豁免以及义务程度的豁免两种角度。《生成式人工智能服务管理暂行办法》第 4 条规定，利用生成式人工智能生成的内容应当真实准确，采取措施防止其生成虚假信息，防止损害他人身心健康,损害肖像权、名誉权和个人隐私，侵犯知识产权。这似乎意味着生成式人工智能对生成的内容负有一般性审查义务，但是这种一般性审查义务又相较于网络侵权规则中的注意义务相对较弱。[③]而在高风险 AI 的具体应用场景，AI 的义务履行程度也应是有限度的，以“医疗”举例，医疗 AI 主要功用是提供诊断结论与推荐治疗方案，并不参与治疗环节，而医疗 AI 的“说明义务的内容说明义务标准，对于在规则构建而言只是提供了一般性概括，但定型化的说明内容才能在裁判规则与行为规则双重意义上提供明确指引。[④]对医疗人工智能的披露以清晰简洁为要，故在众多可能的说明事项中应选取最为核心的部分。[⑤]医疗措施、替代程序以及不采取行动，三者各自的风险和益处是说明义务内容构成的基石，这一通行的告知框架被称为“标准的风险与利益披露”，只要满足“标准的风险与利益披露的”说明义务，对于不合理的说明义务可以拒绝。[⑥]AI 相关主体权益的说明义务在医疗领域的标准始终应该低于真人医生。[⑦]

4.AI 主体的可解释权益

解释性的法教义学的构造其实而并不完整，在“模糊框架”之下，高风险 AI 相关主体应具有解释颗粒度的选择自由。有学者建议，对高风险 AI 的解释可以

---

① 林洹民：《论人工智能立法的基本路径》，载《中国法学》2024 年第 5 期，第 88 页。

② Onitiu, Daria. "The limits of explainability & human oversight in the EU Commission’s proposal for the Regulation on AI-a critical approach focusing on medical diagnostic systems." Information & Communications Technology Law 32.2 (2023): 170-188.

③ 徐伟：《论生成式人工智能服务提供者的法律地位及其责任———以 ＣｈａｔＧＰＴ为例》，载《法律科学（西北政法大学学报）》，第 77 页。

④ 张永亮：《人工智能的场景化应用及其法律规制研究》，浙江大学出版社 2024 年 5 月版，第 32 页。

⑤ Mennella, Ciro, et al. "Ethical and regulatory challenges of AI technologies in healthcare: A narrative review." Heliyon (2024).

⑥ 汪庆华：《人工智能的法律规制路径: 一个框架性讨论》，载《现代法学》2019 年第 2 期，第 59 页。

⑦ Yang, Yujie. "The Duty of Explanation of Using Medical AI." China Law Journal, vol. 2024, no. 1, Spring 2024, pp. 43-58.

采取解释加诠释的方式。基于结构或机制等内部关系的解析直接理解，诠释则是根据其功能或行为等外部关联间接理解。由解释达到的理解更是分析性的，具有由部分说明整体的性质；而由诠释达到的理解则更是整体性的，具有由整体说明部分的性质。但是如果整体诠释需要外部评价时，整体诠释效果应该符合 AI 主体的基本利益，而 AI 主体也有权利进行“解释”的解释。[①]而在“将伦理规范设计 AI 智能系统”的广泛要求之下，AI 的行动、决策是否保证其与设计者的目标一致，或者保证始终与控制者合作，都需要依靠技术本身，但现在的技术并不能保证万无一失，尤其目前机器学习技术并不能确保人机协同的绝对可预定性，更不能保证人机价值观的一致。高风险 AI 相关主体解决运行偏移时，应该有限度的引入可检验、可修正、可控制的人工智能应用模式。[②]法律或者伦理规范的最低要求被满足之后，高风险 AI 相关主体有权利考虑自己运行成本而自己设计方案。这些自由本质上来源于 AI 的商事自由，包括如何处理 AI 运行中的财产性利益。[③]

（四）高风险 AI 的权利义务配置理论

在“如何处理 AI 的侵权”的问题域中，学界研究一直没有跳出传统的侵权在关于“侵权责任主体如何认定？”、“侵权行为如何识别”、“如何界定损害”、“如何判定因果关系”，“如何评价抗辩事由”等主要侵权法类别的研究中，学界的主要思路依然集中在如何让传统的侵权法框架努力适应 AI 技术的发展，很多研究并没有试图打破传统侵权法路径的基本前提，而是在遵从传统侵权法路径基本框架的前提下进行细化和适当的革新。[④]

1.侵权责任主体如何认定？

关于“谁是侵权责任主体”这个问题目前有几个分层的子问题。其一，当 AI 被赋予法律主体资格时，应该如何承担侵权责任？其二，当 AI 和人类设计者同时被认定为责任主体时，又应该如何承担侵权责任。其三，人类设计者的责任。第一个问题，第二个问题暂时不考虑。在第三个问题中，由于 AI 产业链非常复杂，如何在产业链中分配 AI 相关主体的责任？《民法典》侵权责任编关于连带责任的规定可知，立法者往往基于行为人共同实施的活动引起的风险而要求这些行为人承担连带责任，例如《民法典》第 1170 条规定各个行为人承担连带责任，是因为这些行为人实施的行为不仅引起造成损害的风险，而且引起了事实不确定性的风险，第 1214 条规定拼装或者报废机动车的转让人与受让

---

① 王天恩：《人工智能发展中的“智能透明”问题》，载《哲学动态》2023 年第 1 期，第 94 页。
② 龙卫球：《科技法迭代视角下的人工智能立法》，载《法商研究》2020 年第 1 期，第 66 页。
③ Rafael Dean Brown (2021) Property ownership and the legal personhood of artifificial intelligence, Information & Communications Technology Law, 30:2, 208-234.
④ De Cooman, Jerome. (2022). Humpty dumpty and high-risk ai systems: the ratione materiae dimension of the proposal for an eu artificial intelligence act. Market and Competition Law Review, 6(1), 49-88.

人对该机动车引发的交通事故承担连带责任等等。在这些规定中，连带责任并不以参与特定活动的各个主体具有共同故意为限，而是包括共同过失或者一方故意一方过失的情形--只要这些活动引发了统一的风险，本该防范这些风险却未加防范的行为人，就应当对此承担连带责任。[①]AI 系统的风险产生涉及多个主体，如何在这些主体之间配置责任是人工智能监管的一个难题。根据法经济学的思想，为了实现最佳风险防范，责任应该配置给能以最高效率实现最佳目标结果的主体。高风险 AI 人工智能系统使用者是风险管控第一责任人。

2.AI 侵权因果关系的确定

AI 技术让确定侵权因果关系尤为困难。AI 侵权因果关系因不同具体情况而设定不同的条件。比如，AI 只依赖于相关性而学习本身可能就构成错误，因为要素之间的相关性并不是简单的线性关系，相关性是一个网络，而此举可从而穿透内部技术衍变为外部的侵权行为因果关系。[②]再如医疗 AI 未经人类监督进程的过失行为直接导致患者损害结果发生的,其自主决策能力带来的不可预测性将中断医疗机构与患者损害之间的因果关系,在医务人员根据相关手册正确操作智能医疗机器人后,难以认定医疗机构在诊疗活动中未尽到合理注意义务而存在过错。[③]此情形之下可以考虑的追责对象只能是医疗 AI 仪器的生产者（包括程序的开发者）。”[④]对于一般的工具,使用人可以对其实施有效的控制,因此,当在其控制下的工具造成他人损害时,使用人应当承担责任,损害结果与他的控制行为之间存在明确的因果关系,是可以预测的。[⑤]学界研究并不认为需要另辟蹊径不顾虑侵权的因果关系。

3.高风险 AI 的归责原理

针对高风险 AI 人工智能体的应用的主要法律关系涉及双方。既有学术研究关注人们的权利，以期能通过规范性意义的权利义务配置来调整客观上看起来并不平等的关系。通过给弱势方赋权或者给强势方施加义务来源于最朴素的自然法理论。[⑥]AI 的权利配置理论分为共同性的和相对性的。[⑦]共同性的典型代表为利益相关者理论，其不仅强调 AI 相关主体尊重利益相关者的合法权益，也呼

① 阮神裕：《个人信息委托处理中受托人的地位 、 义务与责任》，载《当代法学》2022 年第 5 期，第 118 页。

② 宋珊：《规则: 从界限到相关性———维特根斯坦后期思想对“机器学习”的启示》，载《自然辩证法研究》2022 年第 9 期，第 44 页。

③ Sai, Siva, et al. "Generative ai for transformative healthcare: A comprehensive study of emerging models, applications, case studies and limitations." IEEE Access (2024).

④ 刘建利：《医疗人工智能临床应用的法律挑战及应对》，载《东方法学》2019 年第 5 期，第 138 页。

⑤ 孙山：《论人形机器人的法律地位》，载《东方法学》2024 年第 3 期，

⑥ 丁晓东：《基于信任的自动化决策：算法解释权的原理反思与制度重构》，载《中国法学》2022 年第 1 期，第 114 页。

⑦ Fukuda-Parr, S. and Gibbons, E. (2021), Emerging Consensus on ‘Ethical AI’: Human Rights Critique of Stakeholder Guidelines. Glob Policy, 12: 32-44.

吁利益相关者遵守使用伦理，与 AI 相关主体共负责任、共担风险。[①]而相对性的权利主要赋予 AI 受众，义务则配置给高风险 AI 的相关主体。[②]如 AI 相关主体的答责义务。[③]也有学者认为，在探讨高风险 AI 的运行时，应该早进行“有义务无权利”的非对称性配置。[④]权利义务配置理论还需聚焦配置依据或者配置权威的正当性问题，尤其是高风险 AI 普遍应被受到强硬监管的态度下，维持配置的信服力尤为重要。[⑤]亦有学者认为，在高风险 AI 运用的各个领域，人工主导始终是应用高风险 AI 的黄金原则，亦能够确保 AI 输出结果的正确或至少有人负责。[⑥]可归责的行为表明法律责任或道德责任中归责的一个主要条件，即控制力条件。要满足控制力条件，行动者要对行为有足够的控制力，这样的行为才是行动者自己的行为，而不是被外在因素所完全决定的。在这一意义上，自由意志就是控制力条件的本体论载体。[⑦]

4.产品责任

欧盟新修订的《产品责任指令》已经明确将 AI 产品纳入产品适用范围。产品责任规制范围,除了需要对“产品”作扩大解释外,还需要对 AI 领域的产品责任制度进行改造。讨论人工智能系统的产品责任,关键是界定人工智能系统的生产者，明晰 AI 产品整个生命周期中不同的责任主体所扮演的角色。此外，AI 系统缺陷的概念需要得到明确,在 AI 语境下如何评估 AI 系统是否能按预期安全有效地运行。扩大 AI 损害的赔偿范围也可能有必要。[⑧]通过建立信息披露和提供要求破解 AI 的算法黑箱,并辅以推定等减轻举证责任的规则,在缺陷、因果关系认定等方面适当将天平向弱势的被侵权人倾斜。”[⑨]

5.动物损害责任

高风险 AI 人工智能具有一定的学习能力和较高的自动化程度,可以脱离自然人的控制自行执行任务这的确可能会引起因果关系中断,而难以找到相关方承担相应责任。但是，这种因果关系中断仅仅指的是‘行为’与‘损害后果’的

---

① 张万洪、任文佑：《利益相关者保护的共同善维度——兼评新《公司法》第 20 条》，载《华东政法大学学报》2024 年第 5 期，第 100 页。

② 刘权：《人工智能发展和安全并重的法治探究 ——以人形机器人为例》，载《东方法学》2024 年第 5 期，第 41 页。

③ 刘艳红：《人工智能的可解释性与 AI 的法律责任问题研究》，载《法制与社会发展 (双月刊)》2022 年第 1 期，第 85 页。

④ 周汉华：《论我国人工智能立法的定位》，载《现代法学》2024 年第 5 期，第 32 页。

⑤ Challoumis, Constantinos. "Building A Sustainable Economy-How Ai Can Optimize Resource Allocation." *XVI International Scientific Conference*. 2024.

⑥ 邱遥堃：《机遇与风险：文生视频大模型将如何影响智慧法院建设升级迭代》，载《中国应用法学》2024 年第 2 期，第 98 页。

⑦ 朱振：《归责何以可能：人工智能时代的自由意志与法律责任》，载《比较法研究》2022 年第 1 期，第 64 页。

⑧ Mayrhofer, Ann-Kristin. "Product liability in the age of AI—Proposal for a “two track” solution." (2024).

⑨ 曹建峰、徐艳玲：《医疗领域多模态 AI 模型的机遇、挑战与治理应对》，载《中国医学伦理学》2024 年第 9 期，第 1028 页。

联系中断,而不是‘责任主体’与‘损害后果’的联系中断。由于 AI 能仍旧归属于自然人或法人主体,即便是人工智能自主造成的损害也无法切断‘责任主体’与‘损害后果’之间的联系,此时可以借鉴民法关于“动物致人损害”的责任规则,要求 AI 的所有人承担责任。[①]

在高风险 AI 的应用场景中，损害和风险具有一定的区分，风险性损害本身就是一种尚未确定发生的损害，大部分情形下就应当适用预防性的侵权责任，而非赔偿责任。[②③④⑤]

七、总结

关于学界对高风险人工智能法律治理研究的文献综述表明，该议题的研究价值如下：

第一，不能取缔风险研究进路是 AI 治理研究的基本前提。AI 治理的法律渊源来源广泛，我国网络法体系也基本依照风险进路研究而搭建起来的。在未来的研究中，可以搭配其他路径来配合风险路径的效用，但是风险进路始终是高风险 AI 研究的基本学术背景。

第二，对 AI 不同风险采用不同的治理模式是学界达成的基本共识，也是高风险 AI 研究的问题领域。此问题领域有几个子问题是需要解决而学界却又缺乏关注的。比如，高风险 AI 的识别机制应该如何构建？前述文献表明，单纯依靠分类分级或者应用行业而区别高风险 AI 的法律规范是无法落地，这样会让所有 AI 都可能被判定为高风险 AI，如此，AI 的创新和发展就无从谈起。又如，应该如何看待高风险 AI 与 AI 技术本身的相互关系。既有学术研究对 AI 的定义是具有丰富外延的。技术复杂 AI 并不一定就是高风险的，而“深度合成”这种看起来普遍的技术有可能也会带来严重伤害。上述现实给如何判定高风险 AI 带来了更深层次难题，也使其具有学术研究价值。

其三，高风险 AI 事实上已经运用于各行各业，也带来了不少正面效应，在此基础上高风险 AI 主体运用范畴是值得考量的，学界较少关注 AI 相关主体的权益。如果能为高风险 AI 找寻合适的消化法律路径，高风险 AI 的运行本身是可以接受的，而 AI 主体的正当权益也是值得保护的。

第四，高风险 AI 运用的边界是需要明晰的，学界对既有边界的探讨比较模糊，伦理边界和法律边界并不清晰，如果模糊地带是必然要面对的，那么如

---

① 朱艺浩：《人工智能法律人格论批判及理性应对》，载《法学杂志》2020 年第 3 期，第 139 页。
② 王道发：《个人信息处理者过错推定责任研究》，载《中国法学》2022 年第 5 期，第 119 页。
③ Kan, C.H. (2024). Criminal liability of artificial intelligence from the perspective of criminal law: An evaluation in the context of the general theory of crime and fundamental principles, International Journal of Eurasia Social Sciences (IJOESS), 15(55), 276-313.
④ 郑文革：《人工智能法律主体建构的责任路径》，载《中国应用法学》2022 年第 5 期，第 230 页。
⑤ 夏伟：《论数据犯罪的立法重塑》，载《法制与社会发展 (双月刊)》2023 年第 4 期，第 185 页。

何在模糊地带中切入强制措施的选取角度的规则也很重要。简单来说，高风险AI产生的哪些结果是法律必须要干预的，需要可接受的答案。在公法边界，公法学者对是否要纳入高风险AI入刑还处于如火如荼的讨论中。如果AI涉嫌刑事犯罪，那么应被认为是高风险的，也是人们不能接受的。[①]在边界议题的讨论中，又会出现内部体系的拓展，如如何处理AI法律主体或者刑事责任的问题，如何在刑事规则之外的监管体系中对高风险AI进行包容治理。[②]

第五，如何预防和救济高风险AI也是研究重要部分。欧盟已经颁布了最新的产品指令，希冀把AI纳入产品责任监管机制。表面上，侵权规则似乎具有很多难题，但是如何解构难题以及对侵权法提出怎样的改进还值得进一步研究，比如，是否应该放弃因果关系的基本判定标准或是扩充损害的认定标准，以及如何看待行业内部AI不同的主体责任承担关系。作为严谨行业的AI（金融和医疗领域）运行实践如何指引《民法典》侵权责任编的改进。

第六，以硬法和软法相结合的方式是治理高风险AI的具体路径。既有学术研究对AI的治理路径提出了很多具有建树性建议。现有的很多措施其实并没有站在“不同风险采用不同AI治理”的共识之上，很多学者只是浅尝辄止的提出要注意高风险AI运行，并没有凸显出高风险AI治理措施的特殊性或者必要性。高风险AI的特殊性是其法律治理研究的前提。

[①] [德]埃里克·希尔根多夫，《数字化、人工智能和刑法》，江溯, 刘畅等译，北京大学出版社，2023年12月，第88页。

[②] Kirpichnikov, Danila, et al. "Criminal liability of the artificial intelligence." E3S web of conferences. Vol. 159. EDP Sciences, 2020.

八、参考文献

（一）中文著作

【1】[澳]陈西文：《我们，机器人?人工智能监管及其法律局限》，费秀艳、游传满译，北京大学出版社 2024 年 6 月

【2】[波]多米尼克·埃瓦·哈拉西莫克，《人工智能的伦理与法律》，徐源、刘媛媛译，科学出版社 2024 年 9 月版。

【3】[德]埃里克·希尔根多夫，《数字化、人工智能和刑法》，江溯, 刘畅等译，北京大学出版社，2023 年 12 月。

【4】[美] 瑞恩·卡洛、[美]迈克尔·弗兰金、[加] 伊恩·克尔：《人工智能与法律的对话》，上海人民出版社 2018 年 8 月版。

【5】[美]伍德罗·巴菲尔德，[意]乌戈·帕加洛，《法律与人工智能高级导论》，苏苗罕译，上海人民出版社 2022 版。

【6】[美]伍德罗·巴菲尔德，[意]乌戈·帕加洛：《人工智能法律指南》，冯子轩、汤烨凡译，法律出版社 2024 年 9

【7】[日]弥永真生，宍户常寿，《人工智能与法律的对话 3》，郑超、郭美蓉译，上海人民出版社 2021 年 9 月版。

【8】[英]特纳：《机器人现代法则：如何掌控人工智能》，朱体正译，中国人民大学出版社 2023 年 2 月版。

【9】冯子轩（编），《人工智能与法律》，法律出版社 2022 年 11 月版。

【10】余盛峰：《临界:人工智能时代的全球法变迁》，清华大学出版社 2023 年 5 月版。

【11】岳彩申、侯东德：《人工智能法学研究(第 5 辑)》，法律出版社 2023 年 7 月版。

【12】张永亮：《人工智能的场景化应用及其法律规制研究》，浙江大学出版社 2024 年 5 月版

【13】赵万一、侯东德：《法律的人工智能时代》，法律出版社 2020 年 1 月版。

（二）中文论文

【14】安晋城：《算法透明层次论》，载《法学研究》2023 年第 2 期。

【15】包傲日格乐、曾毅：《信任人工智能何以可能?》，载《自然辩证法研究》2023 年第 2 期。

【16】蔡琳:《算法科技伦理法制化的逻辑证成与建构进路》，载《法律科学（西北政法大学学报）》2024 年第 4 期。

【17】曹建峰、徐艳玲：《医疗领域多模态 AI 模型的机遇、挑战与治理应对》，载《中国医学伦理学》2024 年第 9 期。

【18】常琳、叶凤云：《生成式人工智能用户虚假信息规避行为研究》，载《情报理论与实践》网络首发时间 2024 年 9 月 12 日。

【19】陈昌凤：《智能平台兴起与智能体涌现：大模型将变革社会与文明》，载《新闻界》2024 年第 2 期，第 23 页。

【20】陈全真：《生成式人工智能与平台权力的再中心化》，载《东方法学》2023 年第 3 期。

【21】陈天翔：《算法治理适用比例原则的法理》，载《苏州大学学报（法学版）》2024 年第 3 期。

【22】陈禹衡：《生成式人工智能中个人信息保护的全流程合规体系构建》，载《华东政法大学学报》 2024 年第 2 期，第 49 页。

【23】褚建勋、刘新生：《试论人工智能产品可靠性与企业伦理责任》，载《自然辩证法研究》2022 年第 10 期。

【24】戴昕：《无过错责任与人工智能发展——基于法律经济分析的一个观点》，载《华东政法大学学报》 2024 年第 5 期。

【25】邓建鹏、赵治松：《文生视频类人工智能的风险与三维规制：以 Sora 为视角》，载《新疆师范大学学报（哲学社会科学版）》2024 年第 6 期。

【26】杜严勇：《厌恶算法还是欣赏算法？——人工智能时代的算法认知差异与算法信任建构》，载《哲学分析》2022 年第 3 期。

【27】龚向和、宋凡：《网络平台对数据人格权的尊重义务》，载《南京社会科学》2023 年第 6 期。

【28】顾东晓、黄智勇、朱凯旋、王晓玉、苏凯翔、谢懿：《医疗健康大模型知识体系构建、服务应用与风险协同治理》，载《情报科学》网络首发论文 2024 年 9 月 25 日。

【29】郭剑平：《制度变迁史视域下人工智能法律主体地位的法理诠释》，载《北方法学》2020 年第 6 期。

【30】郭少飞：《算法法人的理论证立及构成要素探析》，载《东方法学》2021 年第 5 期。

【31】郭少飞：《通用人工智能语境下智能法律行为的定位与效力探析》，载《东方法学》2023 年第 5 期。

【32】韩文：《语音数据法律风险防范的本土制度构建》，载《法商研究》2024 年第 5 期。

【33】韩旭至：《生成式人工智能治理的逻辑更新与路径优化———以人机关系为视角》，载《行政法学研究》2023 年第 6 期。

【34】何冠岐、叶茗媛：《涌现系统论———对人工智能“灵性”的呼唤》，载《系统科学学报》2025 年第 2 期。

【35】黄家诚、田广兰：《西方数据正义研究的三条进路》，载《自然辩证法研究》2022 年第 3 期。

【36】黄炜、吴昀璟、余辉、张瑞：《生成式人工智能引发的网络舆情风险预警研——基于引爆点理论分析》，载《情报杂志》网络首发日期 2024 年 9 月 24 页。

【37】黄秀蓉 、赵梓羽：《数字赋能与风险规制之衡平：当环境治理遇见通用人工智能》，载《上海政法学院学报（法治论丛）》2024 年第 5 期。

【38】蓝江：《数字生态下的信息体与智能关联主义》，载《中国人民大学学报》2022 年第 3 期 8。

【39】雷磊：《新科技时代的法学基本范畴：挑战与回应》，载《中国法学》2023 年第 1 期。

【40】李春南、王山：《生成式人工智能时代的语言安全：系统性风险与治理路》，载《国际安全研究》第 2024 年第 5 期。

【41】李彤：《通用人工智能技术提供者义务的审视与优化》，载《东方法学》2023 年第 5 期。

【42】李夏旭：《网络平台间接侵权责任的法理基础与体系展开》，载《比较法研究》2023 年第 3 期。

【43】李亚凝：《无人机分类与法律规制——兼评《无人驾驶航空器飞行管理暂行条例（征求意见稿）》，载《中国应用法学》2019 年第 6 期。

【44】梁远高：《生成式人工智能服务提供者侵权责任的场景化分类及归责认》，载《深圳大学学报（人文社会科学版）》2024 年第 5 期。

【44】凌晨：《有限理性存在者: 智能机器人成为权利主体的理论基石》，载《法学论坛》2024 年第 1 期。

【45】刘建利：《医疗人工智能临床应用的法律挑战及应对》，载《东方法学》2019 年第 5 期。

【46】刘永谋、白英慧：《有限主义视域下的人工智能价值对齐》，载《华中科技大学学报（社会科学版）》2024 年第 5 期。

【47】刘震、崔曦元：《信息传播全过程的人工智能意识形态风险分析》，载《浙江工商大学学报》2024 年第 6 期。

【48】潘雅萍、严火其：《掌控技术：阿米什人与现代技术的关系及其启示》，载《自然辩证法研究》2023 年第 7 期。

【49】齐玥、陈俊廷、秦邵天、杜峰：《通用人工智能时代的人与 AI 信任》，载《心理科学进展》2024 年第 12 期。

【50】商建刚：《生成式人工智能风险治理元规则研究》，载《东方法学》2023 年 3 期。

【51】商建刚：《算法决策损害责任构成的要件分析》，载《东方法学》2022 年第 6 期。

【52】盛学军、邹越：《智能机器人法官: 还有多少可能和不可能》，载《现代法学》2018 年第 4 期。

【53】石佳友、徐靖仪：《医疗人工智能应用的法律挑战及其治理》，载《西北大学学报(哲学社会科学版)》2024 年第 2 期。

【54】宋亚辉：《网络平台的动态规制理论》，载《上海政法学院学报（法治论丛）》2023 年第 2 期。

【55】孙占利：《智能机器人法律人格问题论析》，载《东方法学》2018 年第 3 期。

【56】王策：《智能投顾的法律规制研究》，吉林大学 2024 年博士学位论文。

【57】王高峰、张志领：《算法伦理视域下的脑机接口伦理问题研究》，载《自然辩证法研究》2022 年第 7 期。

【58】王娟、汤书昆：《生成式信息隐私鸿沟伦理风险探讨》，载《科学学研究》2024 年 7 月 19 日。

【59】王利明、丁晓东：《数字时代民法的发展与完善》，载《华东政法大学学报》 2023 年第 2 期。

【60】王利明：《生成式人工智能侵权的法律应对》，载《中国应用法学》2023 年 5 期。

【61】王小伟：《网络中立原则的复合文化批评》，载《哲学分析》2023 年第 2 期。

【62】王莹：《算法侵害责任框架刍议》，载《中国法学》2022 年第 3 期。

【63】王中贝、周荣庭：《新兴技术伦理分析路径研究》，载《自然辩证法研究》022 年第 3 期。

【64】魏斌：《司法人工智能可解释性难题的法律论证分析》，载《法制与社会发展》2024 年第 4 期。

【65】温馨、张超、郭锐、陈凯华、冯泽、朱其罡：《推动我国大模型开源创新生态建设的挑战与建议》，载《中国科学院院刊》2024 年第 8 期。

【66】邬天启：《信息哲学存在论新探》，载《自然辩证法研究》2023 年第 3 期。

【67】夏庆锋：《智能合约的法律性质分析》，载《东方法学》2022 年第 6 期。

【68】徐伟：《生成式人工智能服务提供者侵权归责原则之辨》，载《法制与社会展 2023 年第 3 期。

【69】徐伟：《生成式人工智能服务提供者侵权过错的认定》，载《法学》2024 年 7 期。

【70】许可：《驯服算法：算法治理的历史展开与当代体系》，载《华东政法大学学报》 2022 年第 1 期。

【71】 闫瑞峰：《算法设计伦理治理的立场、争论与对策》，载《自然辩证法通讯》2023 年第 6 期。

【72】杨彪：《协商的代价:数字社会合同自由的认知解释与算法实现》，载《中外法学》2022 年第 2 期。

【73】杨路：《印度人工智能:治理框架、发展现状与趋势》，载《南亚研究季刊》网络首发论文 2024 年 9 月 23 日。

【74】杨显滨：《生成式人工智能服务提供者间接侵权责任的承担与限制》，载法学家》2023 年第 3 期。

【75】叶竹盛：《智能社会中的法治与人的尊严》，载《法律科学(西北政法大学学报)》2023 年第 2 期。

【76】易继明、李春晖：《知识产权的边界：以客体可控性为线索》，载《中国社会科学》2022 年第 4 期。

【77】尹雪萍、王义方：《以风险和责任为核心的算法法律规制》，载《法律适用 2022 年第 10 期。

【78】郁振华：《物自身问题新探》，载《中国社会科学》2023 年第 4 期。

【79】喻国明、曾嘉怡、黄沁雅：《提示工程师：生成式 AI 浪潮下传播生态变局的关键加速器》，载《出版广角》2023 年第 11 期。

【80】喻国明、苏芳：《作为真实世界模拟器的媒介与后真相时代的“拨乱反正”——以 Sora 为例解析数字文明时代的媒介新范式》，载《新疆师范大学学报（哲学社会科学版）》2024 年第 4 期。

【81】袁曾：《生成式人工智能的责任能力研究》，载《东方法学》2023 年第 3。

【82】袁曾：《生成式人工智能责任规制的法律问题研究》，载《法学杂志》2023 年第 4 期。

【83】张海柱：《算法治理中的算法公共性及其实现》，载《自然辩证法通讯》023 年第 6 期。

【84】张洪忠、王彦博、任吴炯、刘绍强：《乌合之众的超级节点？ AI 大模型使用的人机网络结构分析》，载《新闻界》2023 年第 10 期。

【85】张继红：《生成式人工智能生成内容标识义务研究》，载《法商研究》20244 期。

【86】张凌寒：《算法权力的兴起、异化及法律规制》，载《法商研究》2019 年第 4 期。

【87】张凌寒：《中国需要一部怎样的《人工智能法》？———中国人工智能立法的基本逻辑与制度架构》，载《法律科学(西北政法大学学报)》2024 年第 3 期。

【88】张璐：《通用人工智能风险治理与监管初探——Chatgpt 引发的问题与挑战》，载《电子政务》2023 年第 9 期。

【89】张骐：《在技术理性与法治之间———交错时空中的人文主义法学立场》，载《法商研究》2023 年第 6 期。

【90】张守文：《信息权保护的信息法路径》，载《东方法学》2022 年第 4 期，第 0 页。

【91】张素华、李凯：《生成式人工智能虚假信息风险与治理研究》，载《学术探索》2024 年第 7 期。

【92】张涛：《生成式人工智能训练数据集的法律风险与包容审慎规制》，载《比较法研究》 2024 年第 4 期。

【93】张涛：《通过算法审计规制自动化决策以社会技术系统理论为视角》，载《中外法学》2024 年第 1 期。

【94】张夏恒、马妍：《生成式人工智能技术赋能新质生产力涌现：价值意蕴、运机理与实践路径》，载《电子政务》2024 年第 4 期。

【95】张欣：《生成式人工智能的算法治理挑战与治理型监管》，载《现代法学》23 年第 3 期。

【96】张欣：《算法公平的类型构建与制度实现》，载《中外法学》2024年第4期。

【97】张欣：《算法影响评估制度的构建机理与中国方案》，载《法商研究》2021年第2期。

【98】张新平、白朴：《信与不信：通用人工智能信任危机的理论反思与机制再造》，载《内蒙古社会科学》2024年第5期。

【99】张新平：《算法与法律的冲突及其化解》，载《法律科学（西北政法大学学报）》2024年第2期。

【100】张旭芳：《生成式人工智能的算法安全风险及治理路径》，载《江西社会科》2024年第8期。

【101】张玉洁：《重述“技术公益主义”：人工智能立法的一般理论与实践》，载《法治研究》2024年第4期。

【102】郑玉双：《自动驾驶的算法正义与法律责任体系》，载《法制与社会发展 （双月刊)》2022年第4期。

【103】郑志峰：《诊疗人工智能的医疗损害责任》，载《中国法学》2023年第1期。

【104】钟维：《中国式智能投顾：规制路径与方案选择》，载《中国人民大学学报》2020年第3期。

【105】周建军：《生成式人工智能刑事法律风险发生和存在的基本原理》，载《江西社会科学》2024年第8期。

【106】周维栋：《生成式人工智能类案裁判的标准及价值边界》，载《东方法学》2023年第3期。

【107】周学峰：《生成式人工智能侵权责任探析》，载《比较法研究》网络首发日：2024年7月23日。

【108】朱嘉珺：《生成式人工智能虚假有害信息规制的挑战与应对——以chatgpt应用为引》，载《比较法研究》 2023 年第 5 期。

【109】朱荣荣：《生成式人工智能应用中间接识别个人信息的法律保护》，载《科技与法律（中英文）》2024年第4期。

【110】邹昀瑾、牛建华：《新一代人工智能驱动公共安全风险研判！价值向度实践逻辑与路径选择》，载《西安财经大学学报》网络首发时间2024年9月13日。

【111】曾聪：《生成式人工智能与“不可解释性”危害结果的客观归属》，载《华中科技大学学报（社会科学版）》，第60页。

【112】曾庆醒：《涌现效应下的生成式人工智能数据污染及其治理路径》，载《法学杂志》2024 年第 5 期，第 66 页。

【113】蔡悦庭：《算法侵权法律规制研究》，西南政法大学 2022 年博士学位论文。

（三）学位论文

【114】龚琼：弱人工智能发展过程中的法律问题探析，北京邮电大学 2019 年博士学位论文。

【115】贺栩溪：《人工智能算法侵权法律问题研究》，湖南师范大学 2021 年博士学位论文。

【116】于艾思：《人工智能侵权责任制度的法经济学研究》，吉林大学 2022 年博士学位论文。

【117】郑少飞：《人工智能技术应用风险法律分配的制度逻辑》，吉林大学 2023 年博士学院论文。

（四）英文书籍

【118】shoshana zubof, the age ofsurveillance capitalism: the fight for a human fumre at the new frontier ofpower, public affairs press,2019.

【119】winston, jane n,approaches to the study of early human embryo development, tniversity ofcambridge, 1992.

（五）英文析出文献

【120】An Eu Artificial Intelligence Act For Fundamental Rights

【121】Babikian, John. (2024). Balancing Acts: Ethics, Regulation, And Accountability In Ai Law And Policy. 10.13140/Rg.2.2.14654.66888.

【122】Cheng, K., Wu, H. Policy Framework For The Utilization Of Generative Ai. *Crit Care* 28, 128 (2024).

【123】Diane Robinson; Shay Cleary; Sarah Gibson; Andrea L. Miller, "Data Governance And Ai In State Courts," Trends In State Courts 2024 (2024): 9-20.

【124】Hernández-Orallo, J. Evaluation In Artificial Intelligence: From Task-Oriented To Ability-Oriented Measurement. *Artif Intell Rev* 48, 397–447 (2017).

【125】Karen L. Neuman; Dorothy Cory-Wright; Colleen B. Hespeler; Madeleine White, "European Commission's Proposed Regulation On Artificial Intelligence: Conducting A Conformity Assessment For High-Risk Ai - Say What?," Rail: The Journal Of Robotics, Artificial Intelligence & Law 5, No. 2 (March-April

2022): 135-144.

【126】Kate Crawford; Jason Schultz, "Ai Systems As State Actors," Columbia Law Review 119, No. 7 (November 2019): 1941-1972

【127】Mekinsey&Company:What Is Artificial General Intelligence (Agi)?

【128】Quinn P, Malgieri G. The Difficulty Of Defining Sensitive Data—The Concept Of Sensitive Data In The Eu Data Protection Framework. *German Law Journal*. 2021;22(8):1583-1612.

【129】Raso, Filippo, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, And Kim Levin. 2018. Artificial Intelligence & Human Rights: Opportunities & Risks. Berkman Klein Center For Internet & Society Research Publication.

【130】Rawas, S. Ai: The Future Of Humanity. *Discov Artif Intell* 4, 25 (2024).

【131】Scott J. Shackelford; Isak Nti Asare; Rachel Dockery; Anjanette H. Raymond; Alexandra Sergueeva, "Should We Trust A Black Box To Safeguard Human Rights?: A Comparative Analysis Of Ai Governance," Ucla Journal Of International Law And Foreign Affairs 26, No. 1 (Fall/Winter 2022): 35-88.

【132】The Artificial Intelligence (Ai) Global Regulatory Landscape :Policy Trends And Considerations To Build Confidence In Ai.

【133】W. Nicholson Ii Price, "Medical Ai And Contextual Bias," Harvard Journal Of Law & Technology (Harvard Jolt) 33, No. 1 (Fall 2019): 65-116.