科技伦理审查的适应性机制研究

——基于反思平衡理论的分析

Research on Adaptive Mechanisms of Ethical Review of Science and Technology: An Analysis Based on Reflective Equilibrium Theory

马洁 /MA Jie

（苏州科技大学马克思主义学院，江苏苏州，215009）
(School of Marxism, Suzhou University of Science and Technology, Suzhou, Jiangsu, 215009)

摘　要：新兴技术快速发展使传统科技伦理审查面临制度静态性与技术动态性矛盾、价值多元化挑战以及跨学科知识整合困境。反思平衡理论通过在判断、原则与背景理论之间建立动态调适机制，为科技伦理审查提供新的方法论框架。该理论汇集稳定判断并促进持续反思，以夯实科技伦理审查的经验根基。基于具体情境权衡多元道德理论，并引入背景理论以强化知识支撑。在实践层面，通过反思平衡的制度转化、动态调适机制以及多维运行保障体系，推动科技伦理审查从静态合规向动态调适、从规制型治理向反思型治理的范式转换，为应对新兴技术伦理挑战提供了更具适应性和前瞻性的治理路径。

关键词：反思平衡　科技伦理审查　动态调适

Abstract: With the rapid development of emerging technologies, the traditional ethical review of science and technology has been confronted with contradictions between the static nature of the system and the dynamic nature of technology, challenges of value pluralism, and dilemmas in interdisciplinary knowledge integration. Reflective equilibrium theory establishes a dynamic adaptation mechanism among judgments, principles, and background theories, providing a new methodological framework for ethical review of science and technology. This theory consolidates the empirical foundation of ethical review of science and technology by gathering considered judgments and promoting continuous reflection. It weighs multiple moral theories based on specific contexts and introduces background theories to strengthen knowledge support. At the practical level, through the institutional transformation of reflective equilibrium, dynamic adaptation mechanisms, and multi-dimensional operational safeguard systems, it promotes a paradigm shift in ethical review of science and technology from static compliance to dynamic adaptation and from regulatory governance to reflective governance, thereby providing a more adaptive and forward-looking governance path for addressing the ethical challenges of emerging technologies.

Key Words: Reflective equilibrium; Ethical review of science and technology; Dynamic adaptation

中图分类号：B82; O312.2　DOI: 10.15994/j.1000-0763.2026.05.014　CSTR: 32281.14.jdn.2026.05.014

一、问题提出与理论基础

科技伦理审查是指对涉及人的科学技术研究进行伦理方面的外部审查。作为一个制度化的审查机制，它通过应用尊重个人、有益和公

基金项目：江苏省软科学研究项目“江苏人工智能立法的实践路径与风险防范研究”（项目编号：BR2024024）。
收稿日期：2025 年 8 月 27 日
作者简介：马　洁（1984–）女，江苏淮安人，苏州科技大学马克思主义学院讲师，研究方向为伦理学基本问题、科技伦理。Email: maggiedoudou@126.com

正原则来保护研究参与者的权利、安全和福祉。[1]科技伦理审查自20世纪60年代在欧美国家起步以来，已在全球范围内得到推广并发挥重要作用。然而，从基因工程、干细胞研究到人工智能、大数据技术，每一次重大技术变革都在挑战传统伦理审查理论框架的有效性。

1. 科技伦理审查的现实困境

当前科技伦理审查理论框架面临的适应性困境体现在三方面。其一，现有制度框架的静态性与技术发展动态性的矛盾。出于人工智能技术和传统生命科学在技术原理和发展态势的差异，[2]既有的伦理审查框架主要基于传统生物医学研究建立，其相对静态的原则体系难以有效回应新兴技术发展中的动态性、复杂性和不确定性问题。其二，科技伦理审查面临的价值多元化挑战。传统以原则主义为基础的审查办法在应用中被简单化为清单式的操作，在处理伦理原则冲突时可能流于表面的机械平衡。其三，跨学科知识整合存在难点。新兴科技的跨学科特征要求伦理审查能够整合来自科学技术、社会科学、人文学科等多个知识领域的专业判断，诸如“在出版领域应用引发的伦理问题涵盖了数学、哲学、计算机科学、伦理学等诸多学科”，[3]这种整合面临诸多壁垒。

面对科技伦理审查的新问题，部分学者着眼于制度改进，如通过强化自我审查、完善同行审查与推动机构审查等多种途径，[4]优化科技伦理审查实施的组织体系；另一部分学者致力于探讨价值理论建构，分析科技审查伦理原则的价值基础、内容构成，提出在技术设计的源头植入伦理原则的内在进路，[5]抑或讨论具体的新技术语境下伦理原则的适用边界等问题，如自动驾驶科技伦理审查中伦理共识形成机制。[6]尽管现有研究在不同维度上对科技伦理审查制度进行探索，但仍存在一定局限。案例分析研究贴近实践，但难以提炼具有适应性的框架；程序规范的改进无法从根本上解决制度适应性难题。更为根本的是，传统科技伦理审查遵循的原则主义进路，其自上而下的演绎逻辑难以回应技术迭代带来的新情境。因而，亟需探讨更具适应性的伦理审查框架。

2. 反思平衡理论的核心要义及其对不适应问题的启示

本研究引入反思平衡理论作为完善科技伦理审查机制的理论基础。该理论由约翰·罗尔斯（John Rawls）在《正义论》中系统论述。罗尔斯将反思平衡定义为“通过这样的反复来回：有时改正契约环境的条件；有时又撤销我们的判断使之符合原则，我预期最后我们将达到这样一种对原初状态的描述：它既表达了合理的条件；又适合我们深思熟虑的并已及时修正和调整了的判断。这种情况我把它叫做反思的平衡（reflective equilibrium）”。[7]罗尔斯引出反思平衡方法的目的在于“以之为论证秩序良好社会及其成员遵守公平合作条款之可能性的基本方法”。[8]该理论的要义是通过反复调整条件与判断，使原则与深思熟虑的判断达成动态平衡。学界将此定义为“狭义的反思平衡”。与之相比，“广义反思平衡”的概念由诺曼·丹尼尔斯（Norman Daniels）提出，他将背景理论纳入平衡框架，拓展了该方法的理论视野。丹尼尔斯将其描述为“一种试图在特定个人持有的有序三组信念集合中产生一致性的方法，即：(a)一套深思熟虑的道德判断，(b)一套道德原则，以及(c)一套相关的背景理论”。[9]

反思平衡理论对解决科技伦理审查的适应性困境具有重要启示。“‘反思平衡’可帮助我们提出新的公共政策”，[10]它所强调的动态调适为科技伦理审查机制提供了新的方法论基础。伦理审查机制不应固守既定规则，而应在实践中不断修正和完善，以回应新兴科技带来的挑战。基于此，本文立足广义反思平衡理论，系统审视科技伦理审查制度的适应性机制。

二、科技伦理审查的经验维度：深思熟虑的判断的形成

“深思熟虑的判断”是反思平衡理论的认识论起点。在罗尔斯看来，“我们的道德能力最能够不受曲解地体现在这些判断之中”。（[7]，p.37）也就是说，这些判断必须满足特定的条件，才能成为反思平衡的有效输入。在科技伦理审

查语境中，判断不仅包含个体的道德直觉，也涵盖在技术实践中形成的集体性道德经验。正是这些满足条件的判断，构成了科技伦理审查适应性机制的经验基础。

1. 判断形成的认知来源

在科技伦理审查中，稳定判断是指在排除偏见、利益冲突和情绪干扰的情况下经过理性反思形成的道德判断。一方面，这些判断源于多元主体在实践经验与既有原则之间进行的理性反思。审查者在实践中隐性地运用反思平衡的方法，在普遍原则与具体语境间调适，运用情感直觉、想象思维和信任制度等过程达成平衡决策。[11] 另一方面，判断产生于特定情境的道德审议中，包括对技术应用后果的伦理直觉、特定案例的价值评价以及技术发展趋势的价值判断。例如在审查面部识别技术时，审查者对“监控社会”风险的警觉不是抽象的理论推演，而是基于对历史案例和现实应用的客观调查形成判断。在此，对技术趋势的价值判断需具前瞻性。如对人工智能“能力涌现”现象的伦理评估需基于对机器学习发展史的理解，在规模效应、数据依赖和算法优化等多个维度形成判断框架。总之，稳定判断产生于技术应用的具体情境，比抽象原则更能把握科技发展的伦理维度。

2. 多元主体的价值协商

提升科技伦理审查适应性，需要通过多元主体的价值协商达成相对一致的判断。科技伦理审查中的价值协商，一方面需要超越抽象原则的简单应用，在具体情境中判断协调。以自动驾驶汽车的道德决策算法为例，MIT的道德机器（Moral Machine）实验收集了来自全球233个国家和地区超过数百万参与者的4000万个道德决策数据，[12] 揭示了个体生命价值与群体利益的权衡、年龄差异在道德考量中的地位、不同社会群体的保护优先级以及技术效率与道德责任的平衡等道德困境。虽然结果显示“保护生命”这一基本原则得到普遍认同，但在具体的情景中呈现出显著的文化差异和情境依赖性。可见，有效的价值协商不应执着于寻求建立普适的优先级规则，而应容纳多元判断并在具体语境中进行动态调适的理性协商。

另一方面，价值协商的关键在于构建包容性审查机制。传统“专家治理”模式虽具专业判断的权威性，但较为单一的视角难以充分回应技术发展的复杂伦理问题。真正的多元共治不是简单的代表性参与或意见加总，而是不同知识体系和价值立场的实质对话，通过理性论辩达成重叠共识。根据“实践中的集体反思平衡”方法，需要通过整合公众偏好、专家判断和伦理原则的迭代过程来达成情境化的平衡。在合理道德分歧的情况下，公众态度数据在政策制定中发挥重要作用。专家将公众偏好转译为技术参数，阐明不同价值选择的技术后果，促进知情同意的审议。科技伦理审查成为多主体协商平台，技术专家判断可行性，伦理学家阐释道德原则，法律专家界定合规边界，社会科学家分析社会影响，公众贡献基于生活经验的实践智慧。多元价值基础构成了共识的稳定性来源，在保持各自合理内核的前提下寻求共同的原则。

3. 动态持续的判断机制

传统“检查清单”式的伦理审查主要依据既定的伦理原则和标准对技术应用进行一次性判断。该审查模式虽操作简便、标准明确，但在面对新兴技术的快速迭代和不确定性时暴露出局限。以数据伦理审查为例，“现有去识别化制度设计以‘技术处理即豁免义务’为核心逻辑，这意味着数据控制者只需进行一次性的技术处理，便可免除其后续的合规义务。”[13] 该规则仅关注数据处理的瞬时技术状态，而忽视了数据在后续流通中的动态风险。当去识别的数据集在交易、共享、整合过程中与其他数据源结合时，通过关联分析、机器学习等技术，原本匿名的个体身份可能被重新识别。这种“一次评估、永久免责”的模式缺乏对再识别风险的持续监测机制，也未明确数据控制者在全生命周期中的防控义务与责任分配，在数据要素市场化的背景下已无法有效应对数据安全的动态风险。

有效的伦理审查需要在数据全生命周期中进行持续判断。从一次判断到持续判断的转变，

体现了反思平衡理论的洞见。罗尔斯阐述的“反复来回”过程揭示了道德判断的本质特征，即它不是单向的原则应用，而是具体情境中的不断调适和深化的过程。在科技伦理审查中，持续判断意味着将伦理评估嵌入技术发展的全过程，通过动态审视不断校准道德原则与技术实践的平衡，确保伦理判断始终与技术现实保持同步。

由上可知，将深思熟虑的判断作为经验基础，能够揭示技术应用的真实伦理维度。深思熟虑的判断使科技伦理审查超越了抽象原则的机械应用，通过价值协商整合多元道德直觉，在动态审视中保持判断的持续更新，形成扎根具体情境、开放持续调适的特征，为构建适应技术快速迭代的伦理审查框架奠定了坚实的认识基础。

三、科技伦理审查的规范维度：道德原则的层级体系

“反思平衡能够帮助我们从特殊的道德判断得出普遍的道德原则，而且这种普遍的道德原则也具有足以满足道德要求的客观性。”[14]在丹尼尔斯的广义反思平衡框架中，“一套道德原则”指的是用于系统化我们经过深思熟虑的道德判断的一般性规范陈述集合。[9]这些原则处于中等抽象层次，比具体的情境判断更一般，又比哲学背景理论更具体。这正是科技伦理审查适应性机制的关键所在，使其在维持核心价值稳定性的同时，能够灵活应对技术变革。

1. 伦理原则的理论依据

科技伦理审查中存在道义论与目的论两种伦理立场。道义论从行为本身的道德属性出发，将知情同意、隐私保护、人格尊严等视为不可逾越的伦理底线。目的论聚焦技术的社会后果，通过成本效益分析评估技术创新的整体福利影响。两者在科技伦理审查中借助反思平衡方法形成互补关系，以增强对不同情境的适应能力。道义论原则构成审查框架的“硬约束”，确保技术发展不以牺牲人的尊严为代价；目的论考量提供“软导向”，在尊重底线原则的前提下追求社会效益。当二者产生张力时，通过情境化权衡，在保持原则一致性基础上实现有限整合。这种有限整合构成了实现动态平衡的有效策略，即“有限整合论通过将人们关切的道德相关因素纳入到道德决策过程，可以在一定程度上规避人工智能对人类的反向伦理建构。”[15]“有限整合”实施关键在于分层处理，一是底线伦理原则构成技术应用的绝对约束，二是在底线之上的价值冲突允许情境化的权衡，三是在不触及道义底线和核心价值的领域追求效率最大化。

2. 原则体系的柔性框架

科技伦理审查的原则框架是一套用于指导道德判断和规约道德行为的准则体系，它不是单一原则的简单集合，而是多重原则间相互关联的系统整体。传统审查模式倾向于预设原则的优先序列，通过演绎逻辑处理具体问题，但这种静态的层级化处理方式难以应对复杂情境中的原则冲突。反思平衡方法则将原则体系理解为需要在具体情境中不断调适和相互证成的动态结构。

当前科技伦理审查在伦理规范层面依托生命伦理学四原则框架，即自主性、公正性、不伤害和有益性，并纳入透明性、问责性和隐私保护等技术伦理规范，共同构成审查的多重原则体系。这些原则在技术实践的具体语境中不可避免地产生张力。如人工智能系统的透明性要求与商业隐私机密保护存在冲突，算法的公正性与效率优化难以兼顾，认知增强技术是否构成伤害存在争议等。这种冲突源于人类道德价值的多元性以及原则的情境依赖性，即“一切道德价值原则如果离开了语境或处境，因其缺失具体内容、规定、意义的虚空，而不具有现实性”。[16]因此，有效的原则系统不能依赖预设的抽象层级，而必须建立在承认原则间的关联性和相互制约，以及允许原则在具体情境中动态调适的基础之上，在判断与原则、原则与原则之间建立动态的反思平衡，使规范框架既保持内在一致性，又具备回应技术创新的适应性。

3. 原则冲突的动态调和

科技伦理审查面临原则间冲突时，传统二元对立思维或抽象平衡难以有效应对。反思平衡方法提供了动态调和路径，通过情境化分析和分层治理实现原则间的协调。在具体判断层面，需将原则冲突置于特定技术情境中进行评估。例如，基因编辑技术的伦理审查充分展现了分层治理的可能性。世界卫生组织（WHO）专家咨询委员会通过整合科学证据、伦理原则、公众意见和各国经验，形成分层治理框架，包括对体细胞基因治疗采取相对宽松的监管，对生殖系基因编辑实施严格限制，对基因增强保持审慎态度。[17] 原则的权衡在具体情境中得以确定。在反思平衡框架下，分层不是静态的优先级排序，而是根据个体、社会和代际的影响、技术成熟度、科学认知程度和社会共识等动态调整的过程。动态调和的关键在于将原则冲突置于具体技术情境中，综合考量技术成熟度、风险可控性、影响可逆性、社会必要性等多维指标，确定特定情境下的原则优先序，根据技术发展阶段实施差异化治理。

可以看出，情境治理和双向调适是反思平衡方法的核心特征，如学者所言，“反思平衡是这样一种不断调整道德判断和道德原则并使之相互和谐一致的过程”。[14] 这一理论和方法超越了传统的单一理论视角和静态规则体系，使道德原则不再是抽象的教条，而是在与具体判断持续互动中获得更新，为科技伦理审查提供了能够回应技术快速迭代挑战的适应性规范体系。

四、科技伦理审查的知识维度：理论背景的支撑约束

道德原则为科技伦理审查提供了规范框架，但原则的确立和应用必然要嵌入特定知识体系中。丹尼尔斯将其描述为“一套相关的背景理论”，即是一套用于支持道德原则、通过哲学论证揭示不同原则体系优劣的相关理论集合，为评估和选择道德原则提供了超越具体判断匹配的独立依据。背景理论为科技伦理审查的适应性机制提供知识支撑与边界约束。反思平衡理论方法的理论意义正在于将背景理论纳入道德推理体系，使科技伦理审查的适应性建立在知识理性的基础上。

1. 多学科知识的证成支撑

认知封闭是传统伦理审查难以回应科技创新挑战的关键原因之一。广义反思平衡引入背景理论，在科技伦理审查中表现为引入多学科知识体系。从功能看，背景理论拓展了约束原则的判断基础，为伦理原则提供了独立于具体道德判断之外的理论支持，使科技伦理的证成获得了来自多学科的独立论证基础。这种独立证成体现在背景理论能够被一组与约束原则的判断部分不相交的判断所支持。以基因编辑技术为例，当我们评估“应严格限制生殖系基因编辑”这一原则时，需考察人格同一性的哲学理论、遗传决定论的科学认知、代际正义的伦理理论等背景理论的可接受性。这种可接受性既取决于我们对基因编辑本身的道德判断，又受到来自其他领域判断的约束，如教育公平、环境与基因交互的科学判断、气候变化责任等。正是这种部分不相交的判断基础，使背景理论能为道德原则提供独立支持，避免循环论证。此外，背景理论的引入有效解释事实与价值的相互构成，一方面，对技术本质的科学认知界定了伦理评价的可能空间和边界条件；另一方面，伦理关切又引导技术认知的问题意识和研究方向。这种双向关系使科技伦理审查超越单纯的规范演绎，将道德推理深度嵌入技术实践的具体语境中。

2. 跨学科知识的有机整合

背景理论将多学科理论引入道德论证，为伦理原则提供独立于道德判断之外的证成支持，从而满足科技伦理审查的跨学科整合要求。不同学科的认识论基础、方法论范式和价值预设存在根本差异，机械拼接会导致知识碎片化。反思平衡的背景理论为知识整合提供结构框架。以脑机接口技术的伦理审查为例，Neuralink公司2023年获准开展人体临床试验，其审查涉及神经科学、计算机科学、临床医学、伦理学和法学等。技术评估关注设备的生物相容性和信号传输的准确性，医学评估权衡治疗

效益与手术风险，伦理分析需要考虑认知增强对人类本质的影响，法律审查界定脑机数据的所有权和隐私边界。美国食品药品监督管理局（FDA）采用跨学科审查模式建立包含不同领域专家的审查委员会，将技术可行性、风险效益、伦理考量和监管要求纳入统一框架，在迭代对话中形成综合性的审查意见。[18]由此可见，一方面，通过共同指向同一组道德原则，使得各学科知识围绕共同问题形成相互支撑的论证体系。另一方面，通过相互约束实现整合的融贯性。技术可行性为伦理判断提供事实基础，伦理考量反过来约束技术参数设计，法律框架则将两者整合为可操作的规范要求，各学科理论在反思平衡中相互约束和印证，确保逻辑一致性。此外，通过迭代调适实现整合的动态性。当技术评估与伦理关切产生张力时，审查委员会可根据伦理论证调整技术标准，也可根据技术认知重新界定伦理关切的焦点。这种双向调适使得知识整合成为开放的、可修正的过程。

3. 背景理论引导的实质反思

背景理论的引入促使科技伦理审查从静态的程序合规转向实质性的价值反思。传统审查倾向于形式化的合规验证，将复杂的伦理判断简化为标准化的核查清单。而背景理论下的实质反思通过检验道德原则的合理性基础，使每一项审查都成为在特定知识语境下重新审视既定原则的反思过程。背景理论驱动的反思将追问深入到原则的理论根基，而不是停留于程序表面，诸如在脑机接口情境下，当设备可能影响认知功能时，需追问传统意义上的自主是否仍然适用等。此外，背景理论引导的实质反思具有动态性。随着科学认知发展和技术实践经验积累而促使伦理审查原则不断修正。总之，背景理论通过提供独立的理论批判资源、引导情境化分析，将科技伦理审查从“对照清单”的技术操作转化为“理论反思”的实质过程，从而为适应性机制奠定理论基础。

背景理论为科技伦理审查提供不可或缺的知识支撑，与深思熟虑的判断、道德原则共同构成反思平衡视野下提升伦理审查适应性的三个重要方面。总体上看，反思平衡方法为应对技术创新的复杂性和不确定性提供了系统性的方法论工具。

五、科技伦理审查的实践维度：适应性机制的实施保障

理论洞见向实践转化并非自明的过程。因此，探索反思平衡方法向可操作审查机制的转化路径，构建从抽象原则到具体实践的制度化方案，成为推动科技伦理审查范式转型的关键议题。

1. 反思平衡的实践转化

反思平衡方法作为一种道德证成理论，其向科技伦理审查实践的转化，需要解决从认识论框架到操作性机制的转换问题。这种转化要求将抽象的“判断–原则–背景理论”三要素结构转化为可操作、可应用的制度安排。

伦理原则向技术实践的转化包括三种路径。一是价值嵌入路径，即强调通过技术设计将伦理原则物化为技术特征。这不是简单的规则编码，而是“将价值‘写入’技术，运用规范性方法从最初的设计层面缓解技术负效应，建构耦合于技术创新的价值指标”。[19]例如，隐私保护原则需要转化为数据最小化收集、端到端加密、用户控制权等具体技术特征。二是社会技术共构路径，即社会技术系统演化理论主张原则与技术在互动中相互塑造，如“算法公平”概念正是在机器学习实践中逐步明确其操作定义，同时技术发展也推动原则内涵的演进。三是协商诠释路径，通过多主体参与实现原则的情境化转换。在集体反思平衡实践中，公众态度数据作为审议过程的输入，在态度、行为和伦理原则之间建立反思性联系。该机制在综合考虑价值的重要性、利益相关者的偏好以及社会共识程度的基础上，运用层次分析法、多准则分析等决策支持工具，综合专家判断和公众参与来确定价值维度的相对权重。这三种路径在不同技术领域和治理情境中的互补性选择，共同构成反思平衡实践转化的路径体系。

制度化的价值协调框架是实现反思平衡实践转化的核心。基于反思平衡的三要素结

构，科技伦理审查的价值协调包含三个递进层次。第一层次是利益相关者价值判断的收集与分析。通过深思熟虑的判断，在充分信息输入和理性讨论的基础上形成多元价值图谱，为决策提供有效信息。第二层是伦理原则的情境化与权衡。抽象的伦理原则在具体技术语境中转译为可操作的准则。这一过程涉及原则的细化、权重的确定以及适用条件的明确，并根据不同领域和场景确定原则组合与权重配置。第三层是背景理论的整合与应用。聚焦具体决策问题整合跨学科知识。反思平衡围绕“边界对象”完成整合难题，通过风险矩阵、影响评估报告等具体操作为价值判断提供事实基础，避免伦理决策脱离现实。

当然，价值协调框架的有效运作，需要程序正义作为制度性保障。罗尔斯“无知之幕”为审查程序的公正性设计提供理论依据。基于此，科技伦理审查的程序设计中需要涵盖透明议程、包容参与、充分论辩、理由公开、救济途径等程序体系，以确保各方在不预设自身利益的前提下进行价值协商，确保伦理原则能够有效转化为技术规范，并固化为制度安排，从而提升科技治理的公正性和有效性。

2. 动态调适的实施过程

科技伦理审查的动态调适建立在反思平衡的三要素互动机制之上，通过三个相互关联的阶段实现迭代调适。第一，初始平衡的建立阶段。依据深思熟虑的判断、既有伦理原则，结合相关背景理论确立治理原则。这些原则主要涉及人格尊严、基本权利等核心价值，同时整合社会共识形成的规范性要求。第二，实施运用阶段。采用实证方法论系统技术，应用多维度数据，包括技术性能指标、伦理合规状况、社会接受程度等，形成基于证据的效果评估。这一过程产生新的具体判断和经验证据。第三，反思调适阶段。当原则要求与实际后果出现分歧时，启动双向修正程序。一方面，技术实践符合道义却产生负面效应，需重新界定原则的适用边界。如人脸识别的便利性与监控风险的冲突促使我们重新定义合理期待的隐私；另一方面，当后果评估与道德直觉产生冲突，则需重新审视评估框架的完备性并调整价值权重的分配。这种双向校正有利于避免原则的教条和功利计算。这三个阶段不是简单重复，而是螺旋式上升的认识深化过程。每次循环都可能调整对原则的理解，改进评估方法或修正实施方案，使科技伦理治理体系在动态中不断完善。

动态调适在时间维度上贯穿技术发展的全生命周期，体现为过程性治理。前瞻性阶段，在技术研发期开展技术评估，基于现有判断、原则和理论预判伦理风险，形成初步治理框架。过程性阶段，在技术应用期进行跟踪监测，持续收集新的经验证据和具体判断，及时发现原则适用中的问题。回溯性阶段，在技术成熟或出现重大事件后开展效果评价，总结经验教训，系统反思原则和背景理论的适应性，优化治理框架。这种时序性迭代突破了线性因果逻辑，承认技术发展的路径依赖性和突变可能性。尽管这一模式在实践中不可避免的存在问题，“但其实现独立性、权威性的道德伦理判断之逻辑具有一定的借鉴价值”。[20] 此外，动态调适在空间维度上强调普遍性与特殊性的辩证统一。在坚守人类尊严、基本权利等普遍性伦理底线的前提下，承认不同文化、制度和发展阶段对技术治理路径的合理影响。在“同”的层面构成不可妥协的道德底线，在“不同”的层面允许基于国情的差异化选择，实现“和而不同”的治理格局。

3. 持续运行的保障体系

反思平衡机制的有效运行需要建立系统的保障体系，确保审核过程的透明性、参与性和反思性。透明性构成反思平衡有效运作的前提条件。与传统审查的程序透明不同，反思平衡的透明性要求展现三要素互动的完整过程。审查过程需要建立完整的、可追溯的决策链条记录系统，记录涵盖具体判断的收集过程、原则适用的推理过程、背景理论的引入与证成，包括利益相关方的价值主张、冲突焦点的识别过程、协调方案的形成路径以及最终决策的理由阐述等。实施分级信息披露制度，在保护必要商业秘密和个人隐私的前提下，最大程度公开审查标准、程序和结果。正如，提升算法的可

解释性才是推动算法透明目标的可行路径。[21]在技术层面可引入区块链等技术，不仅确保记录的不可篡改性、建立分布式信任机制，而且能为责任追究和经验学习提供可靠保障。

规范化的论辩程序通过制度设计保障实质理性的实现。论证规则的规范化要求参与者基于事实证据和逻辑推理展开论辩，明确区分事实判断与价值判断，避免非理性影响。在判断层面，结构化对话方法有利于达成深层次的价值理解和理性共识，如通过匿名迭代达成共识或者明确分歧点，公众参与机制提供充分审议空间，递进讨论深化理解。在原则层面，建立明确的论证规则，要求对原则冲突的权衡给出充分理由，避免简单的多数决定或权威裁断。同时建立权力制衡机制，通过独立审查、交叉验证等制度安排，防止技术精英主义和利益垄断，维护审查过程的开放性和包容性。

质量保障的核心在于构建基于系统论思维的闭环评估改进机制。[22]构建多层次、多时间尺度的反馈回路是实现持续改进的关键。审核体系需涵盖技术性能、伦理合规、社会接受度、长期影响等维度，通过定期评估、持续监测、事件响应等机制动态捕捉价值整合的实践偏差。同时，明确评估结果与调整行动的对应关系，建立自动触发的改进机制，当关键指标偏离预设阈值时，启动新一轮反思平衡过程。反馈回路的设计需要考虑不同时间尺度，短期反馈关注操作层面的即时问题，中期反馈处理制度层面的适应性调整，长期反馈应对价值层面的范式演进。多尺度反馈的整合推动科技伦理审查从被动应对转向主动适应，形成具有自我调节能力的动态审查机制。

结　语

科技创新的加速演进对传统伦理审查模式提出了深层次挑战。反思平衡理论为科技伦理审查提供了从静态合规向动态调适转型的方法论基础。该理论通过引入深思熟虑的稳定判断、道德原则和背景理论，赋予审查体系必要的适应性和可修正性。在实践层面，通过价值协调的三层框架、动态调适的多维机制以及持续运行的保障体系等制度设计，将抽象的认识论方法转化为可操作的审查程序。在此过程中，“判断–原则–背景理论”三要素在实践中得以实现双向修正。反思平衡理论的应用旨在为构建负责任的科技伦理治理提供理论支撑和实践指引，以期保障科技创新的向善发展。

［参 考 文 献］

[1] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 'The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research'[EB/OL]. Washington: U.S. Department of Health and Human Services, https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html.1979–04–18.

[2] 张平. 人工智能伦理治理研究 [J]. 科技与法律(中英文), 2024,(5) : 1–11.

[3] 徐伟志、潘虹、袁华杰. 人工智能与出版伦理：赋能、挑战与因应策略 [J]. 自然辩证法通讯, 2025, 47（10）: 94–101.

[4] 杜严勇. 人工智能伦理审查：现状、挑战与出路 [J]. 东华大学学报（社会科学版）, 2024, 24（2）: 32–39.

[5] 李伦、徐妍. 人工智能生成内容的实质及其伦理规制路径 [J]. 科技管理研究, 2025, 45（5）: 218–224.

[6] 陈天昊、刘宇尘. 如何治理自动驾驶？——自动驾驶的功能实现、伦理挑战与治理体系 [J]. 浙江学刊, 2025,（4）: 37–48.

[7] 约翰·罗尔斯. 正义论（修订版）[M]. 何怀宏、何包钢等译, 北京：中国社会科学出版社, 2009, 16.

[8] 吴楼平. 反思平衡：罗尔斯理论转向的方法论基础 [J]. 哲学动态, 2020,(3) : 92–100.

[9] Daniels, N. 'Wide Reflective Equilibrium and Theory Acceptance in Ethics'[J]. *The Journal of Philosophy*, 1979, 76(5): 256–282.

[10] 邱仁宗、李亚明. 面向未来的科技伦理治理原则——访邱仁宗研究员 [J]. 哲学动态, 2025,(6) : 155–171.

[11] Morton, J. 'Ethics Review, Reflective Equilibrium and Reflexivity'[J]. *Nursing Ethics*, 2022, 29(1): 49–62.

[12] Awad, E., Dsouza, S., Kim, R., et al., 'The Moral Machine Experiment'[J]. *Nature*, 2018, 563(7729): 59–64.

[13] 郭春镇、林泉君. 用户画像数据去识别化制度之反思与优化 [J]. 浙江社会科学, 2025,(7) : 79–90.

[14] 姚大志. 反思平衡与道德哲学的方法 [J]. 学术月刊, 2011, 43（2）: 48–55.

[15] 孙岩、房帅帅. 有限整合论: 人工智能的道德决策问题新解 [J]. 自然辩证法研究, 2024, 40（9）: 83–90.

[16] 高兆明. 伦理学理论与方法（修订本）[M]. 北京: 人民出版社, 2013, 144.

[17] World Health Organization. 'Human Genome Editing: A Framework for Governance'[R]. Geneva: WHO, 2021.

[18] U.S. Food and Drug Administration. 'Implanted Brain-computer Interface Devices for Patients with Paralysis or Amputation—Non-clinical Testing and Clinical Considerations: Guidance for Industry and Food and Drug Administration Staff'[R]. Washington, DC: U.S. Food and Drug Administration, 2021.

[19] 李洋、梁正. 技术范式的流变: 从效果–效率原则到价值嵌入 [J]. 自然辩证法通讯, 2025, 47（3）: 16–23.

[20] 赵精武. 人工智能科技伦理审查制度的体系化建构 [J]. 当代法学, 2025, 39（1）: 84–96.

[21] 夏明月、陈冬阳. 机器学习伦理风险及其防控治理 [J]. 山东社会科学, 2024,（8）: 106–114.

[22] Cusins, P. 'Understanding Quality Through Systems Thinking'[J]. *The TQM Magazine*, 1994, 6(5): 19–27.

[责任编辑　李斌]