重置风险空间：防范化解生成式人工智能技术风险的认识论研析

Reconstructing Risk Space: An Epistemological Study of Preventing and Resolving the Technical Risks of Generative Artificial Intelligence

范毅强 /FAN Yiqiang[1]　　赵瑞 /ZHAO Rui[2]

（1. 南华大学哲学研究所，湖南衡阳，421200；2. 超威半导体加拿大分公司Markham实验室，安大略，L3R 6G2，加拿大）
(1. Institutes of Philosophy, University of South China, Hengyang, Hunan, 421001;
2. Markham Laboratories, Advanced Micro Devices Canada, Ontario, L3R 6G2, Canada)

摘　要：在农耕文明、工业文明之后，人类已进入信息文明时代。近年，以生成式人工智能（Generative AI）为代表的AI技术触发了新一轮科技革命。本文以技术为视角，在回望GAI技术演化的基础上，发现其技术组织形态包含了多种特征，以结构内嵌与自相似性最为显著，且形成了以Transformer模型为基础的“分形”趋势。基于此，文章从GAI技术风险的逻辑起点出发，还原了技术风险的本质，完成了技术风险的内部证成，澄清了衍生于技术内部的错配、异步和错位风险。并在社会认识论基础上提出了以“重构风险空间”为目标的风险治理对策，防范化解GAI技术风险的生成、叠加与共振。

关键词：生成式人工智能　技术风险　社会认识论　风险治理

Abstract: After the advent of agricultural civilization and industrial civilization, human beings have entered the information civilization era. In recent years, AI technology represented by Generative AI has triggered a new round of scientific and technological revolution. From the perspective of technology, this paper looks back on the evolution of GAI technology and finds that its technology organization form contains a variety of characteristics. Among them, structural embedding and self-similarity are the most significant, which form the “fractal” trend based on the Transformer model. Based on this, starting from the logical starting point of GAI technical risks, this paper restores the essence of technical risks, completes the internal justification of technical risks, and clarifies the mismatch, asynchronous and dislocation risks derived from the GAI technology. On the basis of social epistemology, this paper puts forward technical risk’s countermeasures with the goal of “reconstructing risk space” to prevent and dissolve the generation, superposition and resonance of GAI technical risks.

Key Words: Generative artificial intelligence; Technical risk; Social epistemology; Risk governance

中图分类号：TP18; X820.4　DOI: 10.15994/j.1000–0763.2025.02.013　CSTR: 32281.14.jdn.2025.02.013

随着全球新一轮科技革命和产业变革深入推进，以生成式人工智能（Generative AI，GAI）为代表的AI技术迅猛发展，它彰显人类智慧的同时，重新定义了生产力，改变了人类传统的生产、生活方式。早在2018年，习近平总书记强调，加快发展新一代人工智能是事关我国能否抓住新一轮科技革命和产业变革机遇的战略问题。2023年4月，国家互联网信息办

基金项目：国家社会科学基金一般项目“生成式人工智能的技术风险及其防范化解的哲学研究”（项目编号：23BZX102）。
湖南省社会科学基金重点马院专项一般项目“习近平关于科技伦理重要论述研究”（项目编号：2021YBM16）。
收稿日期：2023年11月29日
作者简介：范毅强（1983–）男，湖南郴州人，南华大学哲学研究所教授，研究方向为科技与社会、社会认识论研究。Email: 24273356@qq.com
赵　瑞（1981–）男，北京人，超威半导体加拿大分公司Markham实验室研究员，研究方向为深度学习研究。Email: hmilyzhao@gmail.com

公室发布了《生成式人工智能服务管理办法》。随着ChatGPT等应用的广泛使用，它宣告了GAI技术已从机器学习时代进入深度学习时代。综观GAI的成熟应用，它们并未包含新的基础技术，而是在通证化、集成化、结构化等全新技术组合下，提升了GAI的易用性和可用性。当人们用GAI憧憬未来时，GAI技术与社会的不可通约性却逐步扩大，如果它脱离以“可解释”“负责任”和“可持续发展”为基础的规制框架，人类历史上从未有过的风险和挑战将会悄然而至。防范化解GAI的技术风险将成为把握新一轮科技革命的战略选择。回溯GAI技术的历史，探知其发展趋势，澄清技术风险的特征和存在形态是本文的目的。

一、生成式人工智能的风险成因与指向

20世纪后期，在“AI能力”与“人类理性”的思辨中，拉开了AI技术风险研究的序幕，也衍生出关于“弱AI”“通用AI”和“强AI”技术的风险分析。[1]依据联合国教科文组织2021年通过的《人工智能伦理问题建议书》可知，推动GAI技术的可持续发展已成为全球共识。作为驶向通用AI或强AI的可能路径，GAI与科技、社会深度交融，关于它风险成因的讨论存在两种侧向。

1. 风险成因的实在论侧向

学者们尝试从逻辑实证主义和科学实在论维度还原GAI的风险成因。以逻辑实证主义为代表的学者认为，GAI理应在实然、应然和似然之间刻画出明确的价值边界方能呈现其可解释原则，似然表现为未来分布的潜在预测。杜严勇等学者认为，GAI在归纳和演绎之间的路径选择并未提供令社会信服的合理解释，AI生成内容是易被操控的。[2]以科学实在论为代表的学者们认为，GAI包含了假定的承诺，且该承诺是从离散的不确定性中生成的。闫宏秀等学者质疑GAI的责任设定。在内容生成过程中，GAI将生成机制、形式和结构进行随意嵌入、延展或缩放，它忽略了价值内部的固有关系。[3]GAI在似然、或然间的差序空间，及其关联的技术黑箱成为风险形成的主因。

2. 风险成因的经验论侧向

经验论支持者在必然性、可能性、偶然性与稳定性的风险推演中，列举了GAI可能引发的乱象，它包括伦理风险、经济风险、政治风险等，阐释了全社会共处于风险渗透过程的危机状态。[4]在GAI算法结构间的时序、差序联结中，学者们反对通证（Token）对意义的绝对表征。认为，通证对意义的建构形式仅类似于前工业社会的熟人信任模式，GAI对意义的重构，会引致社会情感、社会结构的双重失范风险。[5]，[6]由此，GAI算法间的嵌入形式，及其功能结构间的链路组成为反事实条件的形成提供了支撑，如果不控制GAI的使用范围和使用频率，泛滥的AI生成内容会瘫痪社会的调整、反应机制，引发道德失序、伦理失范等社会现象，GAI将成为阻碍社会持续发展的现实困境。

综上，实在论抑或经验论，它们都将技术视为GAI的风险源头，而GAI技术隐含的假定承诺、反事实条件则是技术表层结构下的深层原因。技术风险作为GAI接入人类文明的最大威胁，已在学界成为共识。技术风险（Technical Risk）是指伴随着科技的革新、生产方式的改变而产生的影响技术涉众的预期、要求或取向变化的风险。[7]它既是技术的本质属性，又是技术社会化运行的结果，还是社会文化因素和个人心理特征综合作用的产物。[8]与传统技术风险的治理不同，全新的技术组合、集成方式和结构成为GAI技术风险涌现的主要形式，在GAI社会化运行的过程中，其衍生的真值分配与占先性问题是技术风险涌现的内容构成，形式与内容共同形成了GAI技术的风险空间。因此，从局部或底层结构防控技术风险并不能实现风险防范与化解。场景变化，以及内容生成过程中价值分布与技术意向之间的粘性成为技术风险治理的困境，它表现为GAI的技术结构、技术认知以及技术价值间的冲突。例如，缺少注释的GAI算法，因缺少事实条件的确定性指称，易造成其它开发者的误用。发散的生成空间并未包含与之匹配的技术主体责任，功能嵌

入的多样性又将导致技术主体价值意图的分裂等。防范化解GAI的技术风险需要突破三重困境。

困境1：解释鸿沟。GAI技术结构复杂，技术结构间的聚合关系很难分离，其衍生的意义实在、技术保真又是如此之不透明，且技术演化非但未能消除技术风险，还增加了更多不确定性，解释鸿沟由此形成。将GAI技术的聚合特征、演化趋势与风险嬗变进行深层次关联，有助于揭示技术黑箱；

困境2：责任鸿沟。面对不断演化的GAI技术结构，它不仅具备了嵌入与被嵌入条件，还形成了差异的意向空间，触发风险叠加。可GAI技术背后的责任主体却未能厘清，技术规范的建设条件也一直滞后于技术风险的扩张，责任鸿沟由此形成。澄明各结构间的时序关联，有益于明确技术的责任主体；

困境3：泛化鸿沟。GAI技术缺少有效的场景隔离机制，它可以实现多元场景间的互动与穿梭。GAI内部的技术风险与外部的应用风险之间存在传导路径，容易受到舆论事件或技术缺陷的影响，引发风险共振，泛化鸿沟由此形成。多元场景间的价值认同与价值渗透关系有待进一步澄清。

二、生成式人工智能技术的结构特征与发展趋势

工业革命给予了社会决定性的技术推动，技术与社会的交融空间被逐步分割。技术与社会的时序关联、伴随关联，以及它致使的功能、场景模糊使风险还原成为一项解释工作。澄清技术演化及其关联的风险敞口成为防范化解技术风险的首要任务。20世纪中期，立足符号主义的AI技术在功能回路的向度里进入快车道，尔后，形而上的预设与功能应用间的偏差为技术的演化提供了空间。立足联结主义、生成主义的深度学习开始嵌套进基于符号、功能主义的机器学习，它们构成了新的技术组合方式，并迅速成为AI领域的核心问题，也形成了GAI独有的技术结构。

1. 生成式人工智能技术的非线性特征

20世纪60年代，功能主义驱动下的结构设计并未建构出AI通用模型。学者们愈发感受到，线性的符号操作并不是通达强AI的现实路径，开始从实然与应然间的计算通路转向实然与应然间的似然关系。以计算为主体的算法设计开始由线性功能回路转向非线性预测链路，联结主义开始盛行，它拓展了GAI技术的组合方式。70年代，隐马尔科夫模型以序列为基础，在初始状态层、观测状态层的基础上加入隐藏状态层。至此，开启了计算网络在评估、预测和学习目标下的非线性路径，奠定了非线性网络在概率预测方面的优势。进入80年代，循环、反向传播和卷积神经网络算法相继被提出，它们分别在向量度量、特征提取等方面推动了GAI预测链路的演化，例如，神经网络层级间的非线性关系，呈现了序列间的梯度指向，并将GAI的预测过程等效为神经网络算法中的梯度下降过程。神经网络算法让GAI具备了依据线性或非线性序列度量数据关系的功能，为内容生成提供了条件。

2. 生成式人工智能技术的自相似性与结构嵌套特征

20世纪末，如何解决循环神经网络算法中的梯度消失和梯度爆炸问题成为推动GAI再次演化的关键。长短期记忆网络（LSTM）在循环神经网络的基础上进行了功能性植入，运用“选择性保留或遗忘”功能（门控单元），嵌入了一种类似于记忆的调节器，以摆脱梯度消失或爆炸，也拓展了循环神经网络的捕获长度。2014年，Seq2seq的成功，奠定了GAI的自相似性特征，它立足LSTM在迁移能力差、序列长度受限方面的缺陷，运用两个循环神经网络组成了编码、解码器，构建了“注意力调节”模型，不仅实现了输入序列向目标序列的转换，还提升了模型的迁移能力，建构了GAI全新的技术结构。2017年，Transformer模型在LSTM序列结构的基础上，将编码、解码器提升到6组，组建了多头注意力模型，将Seq2seq中的“注意力调节”发展为“自注意力调节”。2023年3月公布的Chat GPT-4，它已是由12个

Transformer模型块组成的GAI。在实然与应然的空间结构中，GAI技术在自相似性特征下不断提升它的似然函数，Transformer模型已成为GAI技术必不可少的基础构成。

3. 生成式人工智能技术的分形趋势

近年，随着多模态融合技术、时序数据监督、优化方法的提出，GAI获得了新的突破。在多模态任务中，如果多模态时序数据缺少有效监督，梯度下降算法可能被困在一个局部最优中，从而无法获得全局最优解。如何将目标函数的局部最优解转化为全局最优成为推动GAI再演化的新命题。例如，贝叶斯网络可依据局部分布（条件概率分布）描述联合分布，即通过整合特征向量间的局部交互关系，从而获得全局向量关系。GAI技术的分形趋势为GAI从局部最优向全局最优的跨越提供了现实路径。

随着类似于语块（Token）的图块（Patch）方案被提出，跨模态特征向量间的局部交互关系能够在同一制式下被界定，Transformer模型开始突破自然语言模型的范畴，在图像、视频类GAI模型中被使用，并获得了比卷积神经网络更优的效果，这一现象开启了GAI技术的分形趋势。例如，2021年推出的图像GAI模型Vision Transformer，用经典Transformer结构替换了图像GAI模型中的卷积神经网络。[9]2023年10月，Open AI推出基于多模态的GAI分身模型。12月，谷歌发布了基于Transformer 的Gemini模型，再一次奠定了GAI技术的分形趋势。

在技术赋能过程中，以归纳为主的机器学习在多个场域内实现了弱AI，但未能触及强AI。在神经网络嵌套性植入的背景下，GAI在学习迁移、学习泛化方面持续演化。以Transformer为代表的技术结构尝试在确定的信息集合中，结合特征分离、意义分割和分布式计算条件，建立端到端的信任关系。[10]它已成为新一代语言、图像、音频、视频类GAI模型的定式。分形已成为GAI技术发展的阶段性趋势，[11]由Transformer模型映射的技术风险已越来越具象。

三、Transformer模型技术风险的逻辑起点与驱动原因

在“技术风险存在”与“技术风险存在者”双重视域中，技术主体从功能和经验的交互之中单射风险，并以“此在”的方式捕获技术风险。但是，单射的方式远不能厘清技术风险状态与技术嵌套、演化之间的动态关联形式，技术嵌套、演化不仅不会消除存量技术风险，反而会生成增量技术风险，形成新的风险堆栈、分布与扩散形式。Transformer模型在信息加工中呈现出显著的序列状态，它表现为意义的设定、理解或表达和泛化过程。[12]海德格尔提出了理解存在的时间维度，时间使“在世”并从而使“此在”的超越性成为可能。[13]Transformer模型技术风险的存在、存在者、此在于时间维度的超越性从序列方式得以映射。

1. 从意义稀疏到意义稠密序列：错配风险的生成

人类在决策过程中并不遵循贝叶斯法则，而是给予最新经验赋予更多权值，也致使人类在决策时更倚重近期事件。Transformer模型的输入层隐含了此项权重。嵌入（Embedding）机制作为Transformer模型输入层的核心，它将意义稀疏的信息块转换为关联意义信息。它似如伊德描述的具身关系（Embodiment relations），只不过这种意义界定的矢量关系从技术与社会转移到了技术内部。嵌入机制大都采取skip-gram、n-gram算法生成，该算法基于关联、条件概率完成语义设定，统计范围包括中心词周围或前方额定数量的子词，为了降低计算成本、防止过拟和，算法只选择较小的额定数量，不参考边缘概率。显然，Transformer模型被注入了特定的语义设定范式，它依赖语义参数中的向量分布进行意义匹配，也表明Transformer模型并不是依据事实的真实发生概率进行内容生成，该设置为错配风险提供了条件。

错配风险是指，双方或多方及其目标之间的不兼容而导致损失的可能性。Transformer模型的错配问题表现在两方面，一是意义界定

的解释盲区与算力的错配；二是模型的功利倾向与大规模用户的错配。从Transformer模型意义构建的底层算法可得知，虽然GAI的输出内容非常流畅且符合社会公众的认知习惯，但是，合理的意义建构形式并不等于合规、合情的编码方式。在缺少边缘概率的参与下，它的意义设定缺乏合理解释，缺少共识建设。当功利被设定为GAI最重要的权值，结合当下最流行的解答，寻找用户青睐的答案成为终极目标，GAI也运用这一向度迅速获得了用户的认同，用户数量快速增长。但是，它的功利设定脱离了GAI作为社会存在的透明性要求，也构筑了GAI的适配风险，在侵占算力、电力前提下，如果GAI只包含短期逐利的功利选择，缺乏中长期的解释性、负责任、持续发展理念，那么它将与人类文明的发展轨道背道而驰。

2. 从非线性到线性序列：异步风险的形成

虽然Transformer模型存在多种形式，如Encoder-only，Encoder-Decoder，Decoder-only，但是注意力机制始终是Transformer模型的核心组成，其目的是处理过载信息，提高信息的理解或生成效率。参考各类注意力机制，无论是自注意、多头注意力、通道注意力或空间注意力，它们都可以理解为对输入向量的重新规整，以实现对向量的理解或预测。多头注意力可理解为将输入特征拆分、重组成更规整的信息块，并对每一个信息块进行赋值，标注单独的、可训练的权重参数，最后，注意头共用同一个隐藏层输出结果。它类似于伊德所描述的解释关系（Hermeneutic relations），只不过这种意义传导的渗透关系从技术与社会转换到了技术内部。多头注意力机制不能看作一个完整、独立的编解码架构，因为每个注意力头都具备单独的运算设定，但是却共享隐藏层的权重值，这一设定为意义的渗透与整合提供了可能，但却为异步风险的形成制造了条件。

在敏感性、传播时间差异等前提下，因缺乏固定、统一的时间序列，而发生的线程内部或多线程之间的请求、响应或执行冲突，由此引起的风险被称为异步风险。[14]Transformer模型的异步风险主要表现在，模型内部具备多线程加工路径，它不仅存在时间序列的缺失值，而且存在重构时间序列的编辑方法，时间序列的缺场为模型内部责任的划定提供了难题，无法依据风险的形成和变化过程划定责任周期。例如，每个注意力头进行独立的线性变换和注意力计算，最后，多个注意力头的输出通过形状变换进行合并，得到多头注意力的最终输出结果。当效率被设定为GAI最重要的权值，且时间序列缺场的情况下，模型内部的倾向性、意向性将无法估量，风险的渗透、叠加也将无法回溯，也失去了与之对应的责任空间。如果GAI难负责的趋势得以继续，那么它将与人类发展的目标背道而驰。

3. 从绝对时间到相对时间序列：错向风险的拓展

随着片段递归（Segment level recurrence）、相对位置编码（Relative positional encoding）等机制的嵌入，相对时间状态被引入，新一代Transformer模型将特征向量与时间序列之间的对应关系进行了分解和重构，并将它们的差序关系向交叉序列拓展，提出了时间换空间的方案。由此，Transformer模型的处理极限和预测能力大幅提升，它不仅突破了段落与段落之间的语义限制，还保留了更完整的意义关联。甚至，只要显存足够强大，就可以复用更多前端片段（Segment）的隐藏层状态，捕获更丰富的意义向量。如Transformer XL模型等。模型的应用逐渐向各个领域拓展，它们以分形的方式迅速拓展域边界，包括了语音、图像、视频等等。

GAI作为新一代数字生产力，它为数字空间的社会实践和价值秩序提供了新的表征路径。在Transformer模型技术风险的“此在”之维，随着GAI技术的商业推广，技术被各行各业大规模试用，它为感知、行动与决策提供了新的动力。由错配、异步引发的不确定性向社会迅速渗透，而技术实践的逻辑错向又导致技术风险的客观实在性，构成了对技术信任的挑战。它们为风险生成、叠加和共振提供了条件，让难以解释、难以负责的风险周期愈加混乱，学界迫切需要基于技术风险的认识论和方法论

研究，推进显性风险的使然研究，以构建通用的可持续发展框架实现风险周期的穿越。

四、回归社会认识论：生成式人工智能技术风险的认识论探究

回溯GAI的技术风险，技术的不确定性已成为共识，但技术风险的自然化路径依然存在分歧。不同于传统技术，Transformer模型主导的错配、异步和错向风险在GAI技术的社会融合过程中隐射了技术风险的三重结构。不确定的可行性，价值信息的确定性不等价算法结构在价值界定过程中理由的正当性；不确定的合理性，价值信息的确定性和价值界定的正当性并不等同于算法结构责任意图的确定性；不确定的一致性，价值信息、价值界定的确定性，与算法结构责任意图的确定性并不等值于可持续发展理念的一致性。[15]近年，学界呼吁人工智能技术风险与认识论的结合。[16]包括从整合视角阐释科技与人文的关系。[17]从共生维度实施价值选择，并落实于具体实践之中。[18]以及从人与技术关系的视角观察其内在关系。[19]在GAI价值生产的过程中，增量与存量价值发生碰撞，假定承诺与反事实条件的真值边界愈发模糊，在社会文化因素和个人心理特征综合作用前提下，随附于不确定性的风险与敞口也随之变化，社会成为承接GAI技术风险的风险场，基于社会的风险空间成为防范化解技术风险的现实路径。社会认识论为技术治理研究提供了一种认识风险体、理解风险场的网络视角，也为GAI技术风险的认识论研究提供了一种基于社会的网络解释框架。

1. 社会认识论的出场

依始于20世纪中后期，风险本质研究的勃兴与社会认识论研究的勃发是并置的。在社会网络视角下，社会认识论尝试从风险端消解、疏解和化解技术的不确定性，从不确定性的风险体转向风险体与风险场的关联性研究，并确立了技术风险研究的社会认识论视点，应对GAI技术的解释、责任和泛化鸿沟，社会认识论含括了制约、协调与共生三种向度，可作用于错配、异步和错向风险。

制约向度立足于技术的不确定性及其风险敞口，它是呼吁风险防范与规制的前瞻性谋划，以消解不确定性引发的危害为目标。20世纪60至70年代，以数字科技为首的新兴科技率先在人文、社会研究中引发关注，促成了技术风险认识论研究的社会认识论视角。“财富的社会生产伴随着风险的社会生产。”[20]70年代末，人文和社会学者已经开始与工程师一起解决技术的不确定性问题，他们对技术的发展方向进行了宏观和微观分析，提出或优化了风险限制措施，防范技术的社会危害。以增量技术为例，因缺乏确定性规范，制约向度的社会认识论研究主要集中在不确定的可行性维度，学者们阐释的主题以呼吁人文、社会科学与自然科学的合作、新兴科技对传统产业的破坏、新兴科技从业者行为规范的建立，以及新兴科技从业人员应该如何处理道德问题四方面为主。尽管技术的制约向度并未植根于特定的理论，但它从解释、分析视角展示了技术的不确定现象及其本质的危害和风险假设，以及构成这些现象或本质的前瞻知识，为解释鸿沟的消解提供了依据。

协调向度以技术、社会和技术主体为视角，它提倡创新效能与风险规避之间的控制与调节，是以疏解不确定性为目标的前瞻性策划。20世纪末，在现代心理学研究中掀起了一场社会建构主义运动，技术接纳模型与计划行为理论的提出推进了社会认识论的发展，它们从主观态度和规范对技术的社会融合提供了解释。模型基于这样一个事实，主体在风险认同以及风险规避能力上存在差异，技术的不确定性会引发主体的多样化选择。依据不确定性的结构和扩散特征，结合计划行为理论，学者们从技术的存量、增量和潜在用户维度探究技术的社会融合特征。[21]并将技术与主体的态度问题，群体对主体的主观规范问题，转化为信任问题。[22]协调向度下，技术风险被划分为四类，技术风险的生成、叠加和迭代过程中固有的随机性；技术主体决策行为的不可预测和可变性；社会融合过程的不可预测性；以及技术价值延展中出现的意外后果或伴生现象，为责任鸿沟的疏

解提供了指引。

共生向度基于节点和网络特征确立视点，它提倡创新与风险间的调制与解调，认为由相互作用主体组成的系统会自然地向自组织形态发展，[23]它是以化解不确定性为目标的前瞻性规划。21世纪初，无标度网络模型不断丰富，以技术风险扩散为代表的社会网络具备无标度网络特征。[24]网络节点的度数、优先连结特性及其汇聚特征促成了技术风险认识论研究的发展，同质与异质间的渗透、变化与趋势被抽象成网络进行研究。共生向度下，不确定性和风险之间关联的本质是社会性，其网络构成和特征是化解不确定性的基础问题，为泛化鸿沟的化解提供了参考。技术风险感知和风险传递的异质性需要从组织内和组织间的关联进行阐释。创新与风险之间的调制与解调基于技术与群体之间的共生原则，它的目的是负责任地处理风险转化和风险传递的过程，以及创造一种新的动态联结方式提升网络的容错能力，[25]同时防止不确定性增量对网络节点和汇聚状态的破坏。

2. 重构技术风险空间：社会认识论视域下技术风险的防范与化解

GAI技术演化的副现象是不确定性的生成与发展，如果它与社会持续融合，将会重构社会的价值空间，加剧风险的溢出效应。Transformer模型的分形趋势从形而上视角阐释了GAI技术在错配、异步、错向风险中的耦合性。GAI技术的随机性、社会融合的不可还原性和同侪群体的异质性成为解答错配、异步、错向风险的难问题。结合社会认识论在风险治理中的三重向度，社会认识论为解构GAI技术风险的不确定性，还原技术风险的传导路径，构建高效的治理体系提供了理论指引，并从重构技术风险空间的实践路径推动风险治理。

面对GAI技术的不可解释性和复杂的时序关系，社会认识论将风险划定为同质与异质。同质风险指符号体系下，风险主体的可通约性，异质风险是指不能被公因子解释的风险，它与风险的自身特征、社会参与主体紧密相关。个人创造社会结构的同时，也从他人建立的结构中获得反馈，以改变自己和他人的感知方式。[26]GAI技术的功能错配、责任异步、发展错向，以及风险信息的认知偏差和应用错位引发了风险社会中的分歧与认同变化，构筑了差异的耦合结构。例如，GAI技术怀疑论者可能对技术的正向推理持怀疑态度，但却会接受怀疑论者提供的证据。依据社会认识论的风险疏解主张，并非简单的去除由不确定性引发的同质风险，而应正视异质、同质风险间的耦合关系，异质风险间的叠加关系，在安全的解耦条件下寻找疏解风险叠加的可行方案。

当技术风险的非渗透与可渗透被设定为技术主体的特殊性，社会认识论尝试从社会规范与技术意向之间的实质差异疏解风险叠加和共振。风险扩散凝结了共同体内部的规范和态度，在“认同”与“异议”的惯性、封装与补偿之中，形成了与之对应的拓扑结构，并成为风险社会的传输媒介。[27]异质渗透提升了风险叠加的可能性，疏解风险叠加取决于认同体系内部的耦合状态及其拓扑结构的解耦条件。在以风险认同为核心的认知体系中，社会认识论认为，应该阐释异质风险中耦合状态与风险叠加的对应关系，并建立对应的风险疏解理论。当代社会认识论风险叠加研究以质信研究为主，它借助不确定性与确定性之间的分歧差，参考技术主体、技术群体的分布结构和社会的边际期望，尝试对同质化压力与结构性失衡展开调制与解调，化解技术风险中的双向传导效应，降低风险共振概率，防范系统风险。目前，社会认识论视域下防范风险共振现象的措施主要有以下两方面：一是加强同质风险的防范能力；二是改变风险冲击后的尾流场，破坏风险尾流场的一致性传导规律，运用治理策略使异质风险维持在安全常数内。

结　语

就风险社会而言，GAI不只是一种增量技术，它已成为引导社会变革的驱动力，它更是一种影响人类未来的技术。GAI的真正意义在于它作为数字生产力的创新本体预设，及其新的意义界定模式，并基于它的价值功能尝试在

社会中建立新的信任体系，重构社会的价值秩序。本文为GAI风险治理提供了风险介述和方法论引介，从技术内部及其映射的风险形态对技术风险加以阐释。依据GAI技术的结构内嵌、自相似性特征和分形趋势发现，技术内部存在错配、异步和错向风险。基于此，论述了GAI技术在社会融合过程中与不确定的可行性、不确定的合理性、不确定的一致性之间的联动性，形成了由技术和社会组成的风险网络。由此，防范化解GAI的技术风险不仅关乎风险生成，还包含了风险叠加与共振。在创新、风险和责任的价值三角中，社会认识论以保障创新为权值预设，在不确定性延伸与转换的空间中以共生向度应对风险的生成、叠加与共振，以重置风险空间的方式实施风险防范与治理。社会认识论对GAI的政策建议、风险管理和治理策略提供了宝贵的启示。

［参 考 文 献］

[1] 成素梅. 智能社会的变革与展望 [J]. 上海交通大学学报（哲学社会科学版）, 2020, 28（4）: 9–13.

[2] 杜严勇. 智能社会建构中的算法文化: 本质特征、伦理风险及其规避路径 [J]. 同济大学学报（社会科学版）, 2023, 34（1）: 77–87.

[3] 闫宏秀. 负责任人工智能的信任模塑: 从理念到实践 [J]. 云南社会科学, 2023,（4）: 40–49.

[4] 王国豫. 科技伦理治理的三重境界 [J]. 科学学研究, 2023, 41（11）: 1932–1937.

[5] 陈凡、蔡振东. 区块链技术社会化的信任建构与社会调适 [J]. 科学学研究, 2020, 38（12）: 2124–2130.

[6] 李伦、宋强. 技术理性的伦理表征及其超越 [J]. 伦理学研究, 2022,（1）: 92–98.

[7] Ross, R., Pillitteri, V., Graubart, R. et al. 'Developing Cyber Resilient Systems: A Systems Security Engineering Approach Special Publication'[R]. Gaithersburg: National Institute of Standards and Technology, 2021, 69.

[8] 赵建军. 超越“技术理性批判”[J]. 哲学研究, 2006,（5）: 107–113.

[9] Dosovitskiy, A. 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale'[J]. *ArXiv. Org*, 2021, 1–22.

[10] Russell, S., Norvig, P. *Artificial Intelligence: A Modern Approach*[M].New York: Pearson Press, 2022, 315–318.

[11] Yildirim, S. *Mastering Transformers*[M]. Birmingham: Packt Publishing, 2021, 5–7.

[12] Russell, S., Norvig, P. *Artificial Intelligence: A Modern Approach*[M]. New York: Pearson Press, 2022, 315–318.

[13] Heidegger, M. *Sein und Zeit*[M]. Tübingen: Max Niemeyer, 2001, 364.

[14] Kurilin, S. 'Vector Space as an Area of the Operation Risks Characteristics for Asynchronous Electric Machines'[J]. *Journal of Physics Conference Series*, 2019, 1260(5): 1–9.

[15] 范毅强、李侠. 后疫情时代新兴技术道德治理的逆周期路径 [J]. 东北大学学报（社会科学版）, 2021, 23（4）: 1–6.

[16] 肖峰. 人工智能与认识论的哲学互释: 从认知分型到演进逻辑 [J]. 中国社会科学, 2020,（6）: 49–71; 205–206.

[17] 李侠. 基于大数据的算法杀熟现象的政策应对措施 [J]. 中国科技论坛, 2019,（1）: 3–5.

[18] 李建华、吴长静. 后疫情时代的共生伦理及其人类命运共同体实践 [J]. 湖湘论坛, 2020, 33（5）: 74–85.

[19] 易显飞、王克宁. 命物之化而守其宗: 新兴人类增强技术的道家生命文化审思 [J]. 自然辩证法研究, 2023, 39（3）: 3–10.

[20] Beck, U. *Risk Society: Towards a New Moderniy*[M]. London: Sage Publication, 1992, 19.

[21] Moore, G., Benbasat, I. 'Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation'[J]. *Information Systems Research*, 1991, 2(3): 192–222.

[22] McKnight, D., Cummings, L., Chervany, N. 'Initial Trust Formation in New Organizational Relationships'[J]. *The Academy of Management Review*, 1998, 23(3): 473–490.

[23] Koppenjan, J., Klijn, E. *Managing Uncertainties in Networks: A Network Approach to Problem Solving and Decision Making*[M]. New York: Routledge Press, 2004, 6.

[24] Zollman, K. 'The Communication Structure of Epistemic Communities'[J]. *Philosophy of Science*, 2007, 74(5): 574–587.

[25] Grim, P. 'Threshold Phenomena in Epistemic Networks'[R]. AAAI Fall Symposium-Technical Reports, Washington, United States, 2009, 53–60.

[26] Burke, P. *Contemporary Social Psychological Theories*[M]. California: Stanford University Press, 2006, 88.

[27] 范毅强、鲁世林、李侠. 新兴技术认同融合中的渗透现象及其实证研究 [J]. 自然辩证法通讯, 2022, 44（7）: 34–44.

［责任编辑　李斌］