农业图书情报学报
DOI：10.13998/j.cnki.issn1002-1248.26-0594

从共同风险到治理合作：中美人工智能安全治理的政策议题结构与合作路径

蔚海燕[1]，崔文波[1,2*]
（1. 华东师范大学 经济与管理学院，上海 200062；2. 华东师范大学 国家教育宏观政策研究院，上海 200062）

摘　要：[目的/意义]人工智能安全风险具有跨境外溢性，中美治理选择对全球人工智能安全治理具有重要影响。本文旨在识别两国安全治理的政策议题结构及跨国关系，并据此分析竞争条件下的合作空间。[方法/过程]基于2020—2026年中美50份人工智能治理政策，经文本筛选与预处理提取4 188条治理命题，采用BGE-M3与BERTopic开展联合主题识别，并结合文档加权关注度及MIX、BAL、CAI指标进行跨国比较。[结果/结论]研究识别出6类治理主题。AI风险识别、评估与全周期治理以及数据安全、隐私保护与权利保障具有较高的双边关注程度；标准化、测试与评估存在一定治理工具交集，但政策权重不对称；透明披露与合规责任、治理能力建设以及国家安全导向的AI基础设施呈现较明显的政策重心差异。中美合作可优先围绕共同风险开展专业交流，以标准、测试与评估作为技术协调接口，并根据跨境外溢性、治理工具可协调性和战略敏感程度确定合作深度，由政府间沟通、科学专家网络、标准技术组织及多边和二轨机制承担相应功能。研究表明，中美人工智能安全治理同时存在共同风险关注、有限技术交集和政策重心分化，合作安排需结合具体议题属性进行差异化配置。

关键词：人工智能安全；中美关系；政策文本分析；治理合作；全球人工智能治理
中图分类号：G203；TP18　　**文献标识码：**A　　**文章编号：**1002-1248（2026）10-0004-11

引用本文：蔚海燕，崔文波．从共同风险到治理合作：中美人工智能安全治理的政策议题结构与合作路径[J]．农业图书情报学报，2026，38(10)：4-14.

0　引　言

人工智能技术加速迭代及其应用范围持续扩展，引发的相关风险已由算法偏见、隐私侵权和信息失真延伸至模型失控、恶意使用、网络攻击、生物安全和军事应用等领域[1]。人工智能风险具有跨境传播、技术扩散及潜在系统性影响，其治理逐渐超出单一国家的政策边界[2]。构建全球人工智能治理体系[3]，加强风险识别、安全评估及治理规则协调，已成为全球人工智能治理的重要议题。作为人工智能研发和应用的重要国家，中美两国的治理选择不仅影响各自人工智能发展，也对全球人工智能安全治理产生重要影响[4]。例如，在2023年11月，中美共同参与《布莱切利宣言》等国际人工智能安全治理进程[5]，并于2024年启动人工智能政府间对话，就人工智能风险和全球治理等问题开展沟通[6]。但中美在发展模式、治理理念和

收稿日期：2026-08-28
基金项目：国家社会科学基金一般项目“数据重用视角下的研究型图书馆科学数据服务模式研究”（22BTQ020）
作者简介：蔚海燕（1980- ），博士，副教授，博士生导师，研究方向为数据管理与服务
*通信作者：崔文波（1998- ），博士研究生，研究方向为科研数据治理。E-mail：52284419012@stu.ecnu.edu.cn

政策工具等方面仍存在差异，科技竞争与国家安全考量增加了治理协调的复杂性[7]。如何在合作需求与现实约束并存的条件下推进治理协调，成为亟须解决的问题。

人工智能治理涵盖产业发展、创新促进、公共部门应用和风险规制等多类政策内容，本文将其中围绕人工智能研发与应用安全风险形成的风险识别、评估、控制、责任配置和制度协调安排界定为人工智能安全治理，重点考察与风险识别、安全保障及治理响应相关的政策内容。围绕这一政策领域，现有研究较多从国际竞争、治理理念、规则差异及具体合作事件等方面讨论中美人工智能治理关系，但对两国安全治理政策形成了哪些议题、不同议题在各自政策体系中处于何种位置，以及这些议题关系如何影响合作空间，仍缺乏系统分析。

基于此，本文以2020—2026年中美人工智能治理政策为研究对象，通过联合主题识别与跨国比较，考察两国人工智能安全治理的政策议题结构，即不同治理议题在政策体系中的分布、相对权重及跨国关联。本文重点回答两个问题：中美人工智能安全治理政策形成了怎样的议题结构，不同议题在两国政策体系中的相对位置和关联程度如何？如何依据不同议题的治理属性与现实约束形成差异化合作路径？本文从政策议题层面分析共同风险向治理合作转化的现实空间及其实现方式。

1　文献综述

1.1　人工智能伦理与安全治理研究

人工智能治理研究早期主要围绕伦理原则及其制度化展开。现有研究普遍将透明、公平、责任、隐私和不伤害等作为人工智能治理的重要规范内容[8-10]。随着人工智能能力提升及应用范围扩展，研究重点逐渐延伸至恶意使用、网络攻击、生物风险、自主系统失控等高影响安全风险，并关注风险识别、能力评估、风险控制和治理响应等环节[11]。相关政策比较研究也表明，安全、隐私和数据保护已成为主要治理主体共同关注的重要内容，但不同国家在风险来源、责任配置和政策工具方面存在明显差异[12]。现有研究已经形成从伦理规范到安全风险治理的基本分析脉络，为识别人工智能安全治理的主要政策内容提供了理论基础。但相关研究较多围绕风险类型、伦理原则或单项治理措施展开，对不同安全治理议题在国家政策体系中的相对位置及其组合关系关注不足。

1.2　中美人工智能治理比较研究

中美人工智能治理比较主要集中于治理理念、国家利益、制度结构和政策工具等方面。已有研究从宏观层面揭示了中美及其他主要治理主体在利益目标、治理结构和政策工具配置上的差异，并将其与战略竞争、“国家-市场关系”和制度环境联系起来[7]。近年来，相关研究开始转向具体风险和治理工具，通过比较中美政府及企业治理文件，识别系统可靠性、危险能力评估及部分安全治理措施等具有较高内容重合度的议题[13]。该研究进展推动中美人工智能治理比较由总体模式分析进入具体政策内容比较，但仍存在两个不足：一是已有研究较少在统一分析框架下考察不同安全治理议题的整体结构及其相对权重；二是现有比较研究对政策内容的重合已有关注，但对宏观制度差异、具体议题关联及其合作条件之间的联系仍缺乏系统讨论。

1.3　人工智能安全治理的国际合作研究

人工智能安全治理合作研究主要关注国际制度安排、风险属性和战略约束。相关研究指出，不同国际治理架构在协调效率、参与范围、制度弹性和治理能力方面存在差异，国家利益分歧和治理能力差距也会影响全球人工智能治理的协调效果[14-16]。从风险属性看，不同类型的人工智能风险具有不同的跨境影响和治理需求，需要匹配相应的国际公共产品供给方式及治理层级[17]。双重用途技术研究则表明，前沿人工智能治理还受到核查能力、透明度、战略竞争、安全保护和规则执行等因素制约[18]。现有研究说明，人工智能安全合作不能脱离具体风险属性和战略环境进行讨论，但对不同政策议题应采用何种合作深度、技术协调方式和合作机制，仍缺乏基于中美政策结构的系统分析。

综上所述，现有研究已从安全风险、中美治理比较和国际合作等方面形成较为丰富的讨论，但对中美人工智能安全治理议题的整体结构、不同议题的跨国

DOI：10.13998/j.cnki.issn1002-1248.26-0594

关联及其合作条件仍缺乏系统分析。基于此，本文从政策议题层面开展联合主题识别与跨国比较，并结合议题属性讨论差异化合作路径。

2 研究设计

2.1 数据来源与样本构建

本文以人工智能治理与监管档案（AI GOvernance and Regulatory Archive，AGORA）为基础构建中美人工智能治理政策语料。AGORA对人工智能相关法律、行政政策、治理框架和技术规范进行结构化收录，并提供政策文本及片段级标注[19]。AGORA对美国政策的覆盖程度高于中国，为补充中国政策样本，本文从中国政府及国家级治理机构官方网站获取相关政策文件。

样本筛选以2020年至2026年8月20日的中央或联邦层面正式政策为范围，纳入实质涉及人工智能发展、安全、伦理、风险或治理安排的政策文件，排除重复文本、非正式评论及缺乏实质政策内容的材料。依据上述标准，最终形成中国22份、美国28份政策文件，共50份政策样本。主题分析聚焦能够形成明确治理分析单元的政策内容，即文本中包含可识别的治理原则、风险对象、规则工具、治理措施或实施要求。按照统一标准，经人工复核后，最终43份政策进入主题模型，其中中国15份、美国28份；其余战略背景类文件保留用于政策语境分析。

本文以治理命题作为主题分析的基本单位，指能够相对独立地表达一项主要风险、治理原则、规则要求或政策行动的语义单元。例如，经文本处理后，中国《新一代人工智能伦理规范》[20]中的“充分尊重个人信息知情、同意等权利，依照合法、正当、必要和诚信原则处理个人信息，保障个人隐私与数据安全”构成一条中文治理命题；美国*Blueprint for an AI Bill of Rights*[21]中的“You should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used”构成一条英文治理命题。两者均能够相对独立地表达明确的治理要求。命题切分采用确定性文本规则完成：在规范化文本和清除版式噪声后，优先依据列表结构及中英文句末标记进行切分，对包含多个治理行动的长文本，根据治理动作和主体结构进行细分，结合治理动作词、治理对象词及语句完整性规则筛选有效命题。每条命题均保留来源文件、原始片段、文本位置及切分规则，以维持其与政策原文的对应关系。对治理命题的语义完整性及其与原文的对应关系进行人工复核后，最终获得4 188条有效治理命题，其中中国1 322条、美国2 866条。

2.2 文本处理与语义表示

文本预处理遵循最小干预原则。首先进行Unicode规范化、控制字符和版式标记清理，删除目录、页码、重复标题等非实质性内容。针对预实验中出现的语言与国家来源分离现象，本文依据文本字符构成识别实际语言，不再以政策所属国家作为语言判断依据。最终，中国治理命题中773条识别为中文、549条识别为英文，美国2 866条命题均识别为英文。

为使中英文政策进入统一语义空间，中文治理命题采用Helsinki-NLP/opus-mt-zh-en模型转换为英文，原本以英文收录的政策文本则直接保留原文。对于较长文本，按模型可处理长度进行分段转换，并采用束搜索生成英文文本。原始文本与英文语义表示保持一一对应。在语义向量构建方面，采用BGE-M3模型生成1 024维文本嵌入，并进行L2归一化[22]。

2.3 联合主题识别

主题识别采用BERTopic框架[23]，通过语义嵌入、UMAP降维和HDBSCAN密度聚类识别政策文本中的潜在治理主题。HDBSCAN依据样本密度形成聚类，不预先设定主题数量。考虑到中美治理命题数量及来源文档分布存在差异，本文采用“国家-文档平衡”的主题发现策略，每次分别在不同来源文档间轮流、无放回抽取700条治理命题，并设置20个随机种子重复运行。

为确定最终模型，本文比较不同重复运行之间的聚类一致性，在两两运行共同抽取的命题上计算调整兰德指数（Adjusted Rand Index，ARI）和标准化互信息（Normalized Mutual Information，NMI），并以各次运行相对于其他运行的平均ARI作为主要选择依据。20次运行形成5~8个主题，其中编号10的运行结果平均ARI最高（0.608），确定为最终模型，并形成6个

一级治理主题。主题层面的Jaccard相似度用于考察不同重复运行之间的主题重现性；关键词、代表性治理命题及来源政策用于最终主题的人工审核和命名。

选定模型后，将全部4 188条有效治理命题映射至统一主题空间。鉴于不同主题内部的语义集中程度存在差异，本文依据主题发现阶段归入各主题的命题与相应主题中心之间余弦相似度的经验分布设置主题特定归属阈值[24]，以第5百分位作为最低归属界限。最终3 901条治理命题获得有效主题归属，287条保留为未归类文本，占全部命题的6.85%。最终聚类结果如表1所示。其中，T2在中国政策命题中未形成对应的独立主题聚类，因此中国侧关键词记为“—”；该结果仅表明相关内容在当前样本中未形成独立语义集群，并不代表中国政策完全缺少相关议题。

表1　中美人工智能安全治理政策主题识别与人工命名结果

Table 1　Topic identification and manual labeling results for China-U. S. AI safety governance policies

Topic	主题名称	中国政策Top10关键词	美国政策Top10关键词	主题内涵
T0	透明披露、内容标识与合规责任	服务提供者；安全；法律；用户；算法；算法推荐；气象；拟人化交互；推荐；交互	human; registration; material; work; copyright; authorship; applicants; oversight; automated; ai-generated	围绕AI活动中的信息披露、内容标识、用户告知、备案监督和主体责任，强化AI应用的透明性、可识别性与责任可追溯性
T1	标准化、测试与评估	标准；安全；规范；研制；语音；标准研制；计算；框架；基础；测试	standards; training; research; performance; technical; software; automated; developing; benchmarks; testing	通过技术标准、测试方法、性能指标和评价基准等形成AI技术与安全治理的共同尺度
T2	国家安全导向的AI基础设施与联邦协调	—	security; infrastructure; energy; sites; consultation; committee; frontier; defense; standards; law	围绕前沿AI基础设施、能源与算力保障、国防安全及联邦跨部门协调，构建国家安全导向的治理体系
T3	数据安全、隐私保护与权利保障	安全；风险；个人信息；权益；权利；保护；影响；隐私；知识产权；安全风险	privacy; rights; sensitive; access; civil; risks; security; domains; safety; protections	重点规范AI开发和应用中的数据处理活动，并保护个人隐私、敏感信息及相关基本权利
T4	治理能力建设、创新应用与多方协作	治理；安全；安全治理；框架；全球；治理框架；全球治理；共识；合作；国际	security; research; risk; infrastructure; responsible; governance; innovation; framework; training; practices	通过制度、组织、人才、研发以及国内外多主体协同，提高AI发展应用与安全治理能力
T5	AI风险识别、评估与全周期治理	安全；伦理；风险；科技；科技伦理；治理；应用；社会；安全治理；影响	risks; automated; human; research; algorithmic; rights; safety; bias; potential; privacy	围绕AI生命周期开展风险识别、评估、审查、监测、人工监督与风险缓释，实现持续性风险治理

如表1所示，中美人工智能安全治理政策主要形成透明披露与合规责任、标准化与测试评估、国家安全导向的AI基础设施、数据与隐私保护、治理能力与多方协作以及AI风险全周期治理6类主题，覆盖技术规范、风险控制、权利保障、治理能力和战略安全等不同层面的政策议题。其中，T2仅在美国政策命题中形成独立主题聚类，而其余主题均包含不同程度的中美政策内容。

2.4　跨国主题比较测度

主题规模及两国治理命题数量可以反映聚类的基本构成，但难以直接表示不同主题在两国政策体系中的相对关注程度。为从不同层面刻画中美人工智能安全治理主题的跨国关系，本文设置主题内部混合度（MIX）、双边平衡程度（BAL）和共同关注强度（CAI）3个指标，分别考察同一主题中是否同时包含两国政策内容、两国对该主题的相对关注程度是否接近，以及双方是否均对该主题保持较高关注。

设$N_{CN,t}$和$N_{US,t}$分别为主题t中中国和美国治理命题数量，则主题内部混合度（MIX）定义为：

$$MIX_t = \frac{2\min\left(N_{CN,t}, N_{US,t}\right)}{N_{CN,t} + N_{US,t}} \tag{1}$$

MIX反映同一主题在命题层面的跨国共现程度。当某一主题完全由一国政策命题构成时，MIX取0；两国治理命题数量越接近，MIX越接近1。该指标主要描述主题内部的双边构成，仍会受到政策篇幅和治

DOI：10.13998/j.cnki.issn1002-1248.26-0594

理命题数量差异的影响。

为比较主题在两国政策体系中的相对位置，本文采用文档等权方式计算国家层面的主题关注度。设国家 c 实际进入主题模型的政策文件数量为 D_c，政策文件 d 中归入主题 t 的治理命题数量为 $N_{d,t}$，该文件全部有效治理命题数量为 N_d，则国家 c 对主题 t 的文档加权关注度为：

$$P_{c,t} = \frac{1}{D_c}\sum_{d \in c}\frac{N_{d,t}}{N_d} \tag{2}$$

其中，$c \in \{CN,US\}$。未归类命题仍保留在 N_d 的分母中，以保持不同政策文件主题占比计算口径一致。文档等权处理赋予每份政策文件相同权重，可以降低少数篇幅较长政策对国家主题关注结构的影响。

基于两国的文档加权关注度，主题 t 的双边平衡程度（BAL）定义为：

$$BAL_t = \frac{2\min\left(P_{CN,t}, P_{US,t}\right)}{P_{CN,t} + P_{US,t}} \tag{3}$$

BAL 用于衡量中美两国对同一主题的相对关注程度是否接近。当某一主题仅在一国形成政策关注时，BAL 取 0；两国主题关注度越接近，BAL 越接近 1。该指标刻画的是同一主题在两国政策结构中的相对平衡程度。

共同关注强度（CAI）为：

$$CAI_t = \frac{2P_{CN,t}P_{US,t}}{P_{CN,t} + P_{US,t}} \tag{4}$$

CAI 用于衡量双方对同一主题的共同关注水平。只有当两国对该主题的文档加权关注度均较高时，CAI 才会取得较高数值；任一方关注度较低都会限制该指标水平。CAI 越高，说明两国均在政策文本中对该主题保持较高关注。

3　中美人工智能安全治理的政策议题结构与跨国比较

3.1　中美人工智能安全治理的主题关注结构

在联合主题识别基础上，本文采用文档加权关注度比较中美人工智能安全治理的主题结构。与治理命题绝对数量相比，文档加权指标赋予每份政策文件相同权重，可以降低少数长文本对国家主题结构的影响。各主题的文档加权关注度及跨国比较结果（表2）。

表2　中美人工智能安全治理主题的跨国比较结果

Table 2　Cross-national comparison of AI safety governance topics in China and the United States

Topic	主题名称	P_{CN}	P_{US}	BAL	MIX	CAI
T0	透明披露、内容标识与合规责任	0.388	0.020	0.099	0.324	0.038
T1	标准化、测试与评估	0.165	0.032	0.325	0.597	0.054
T2	国家安全导向的AI基础设施与联邦协调	0.000	0.127	0.000	0.000	0.000
T3	数据安全、隐私保护与权利保障	0.104	0.120	0.929	0.597	0.112
T4	治理能力建设、创新应用与多方协作	0.021	0.367	0.108	0.128	0.040
T5	AI风险识别、评估与全周期治理	0.242	0.195	0.893	0.713	0.216

如图1所示，中美人工智能安全治理在主题关注结构上呈现明显差异。中国政策关注度最高的是透明披露、内容标识与合规责任（T0，0.388），其次为AI风险识别、评估与全周期治理（T5，0.242）和标准化、测试与评估（T1，0.165）。美国政策主要集中于治理能力建设、创新应用与多方协作（T4，0.367），其次为AI风险识别、评估与全周期治理（T5，0.195）、国家安全导向的AI基础设施与联邦协调（T2，0.127）以及数据安全、隐私保护与权利保障（T3，0.120）。

MIX 和 BAL 分别从议题内部的跨国混合程度和两国相对关注程度两个维度刻画主题的跨国关系。MIX 较高，说明同一主题中来自中美两国的治理命题均有较充分分布；BAL 较高，则说明两国对该主题的政策关注程度较为接近。二者结合后，可以区分3种主要情形：当MIX和BAL均处于相对较高水平时，表明双方在同一议题上既具有较明显的共同进入，又保持较接近的关注程度；当MIX较高而BAL相对较低时，说明双方在治理内容或工具上存在一定交集，但政策关

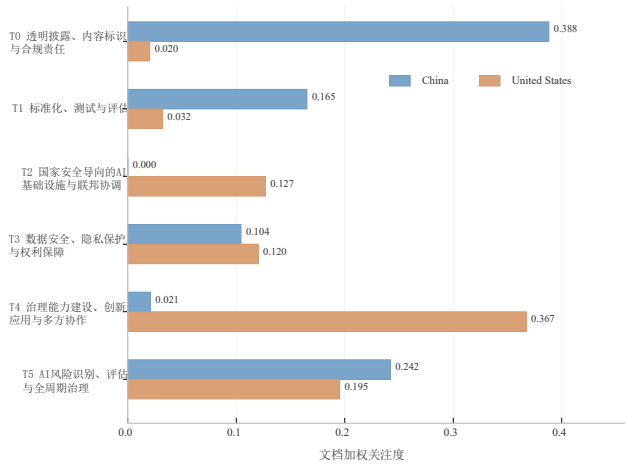


图1　中美人工智能安全治理主题的文档加权关注度

Fig.1　Document-weighted attention to AI safety governance topics in China and the United States

注权重并不对称；当MIX和BAL中至少一项处于较低水平时，则说明该主题更多表现为单方集中关注或跨国分布不足。基于此，本文结合各主题在二维空间中的相对位置及其具体政策内容，对6类主题的跨国关系进行归纳。如图2所示，T3和T5主要表现为共同关注，T1主要表现为治理工具层面的技术交集及关注不对称，T0、T2和T4则更多体现政策重心和治理功能上的结构性分化。

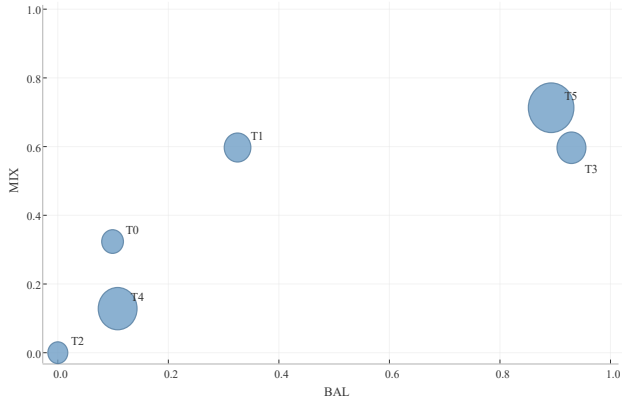


图2　中美人工智能安全治理主题的跨国关系特征

Fig.2　Cross-national relationship patterns of AI safety governance topics in China and the United States

3.2　风险治理与数据权利的共同关注

在6个主题中，T5和T3的双边共同关注程度相对较高。T5的BAL为0.893、MIX为0.713，CAI达到0.216，为全部主题最高；T3的BAL达到0.929，为6个主题最高，CAI为0.112。两类主题均同时进入中美政策体系，并具有较为接近的相对权重。

T5主要涉及AI风险识别、评估、监测和持续控制。中国政策较多涉及风险分类分级、科技伦理审查、风险监测预警、人工复核和风险处置，《人工智能安全治理框架2.0》[25]和《人工智能应用伦理安全指引1.0》[26]均将风险识别和持续治理贯穿人工智能研发与应用过程。美国相关政策更多涉及风险评估、测试验证、审计、持续监测和人工监督，*National Artificial Intelligence Research and Development Strategic Plan 2023 Update*[27]及国家安全领域相关治理文件均强调对技术及社会技术风险的持续识别和缓释。中美两国在具体工具配置上存在差异，但均覆盖事前评估、过程监测和风险处置等关键环节。T5的共同关注主要体现为对持续风险治理目标的共同关注。

T3呈现相似特征。中国政策重点涉及数据安全、个人信息保护、访问权限和隐私保护，美国政策则集中于隐私、敏感数据、知情同意、数据访问控制及公民权利保护，*Blueprint for an AI Bill of Rights*集中体现了相关政策要求[21]。双方在权利表述和制度安排上存在差别，但均将数据处理过程中的安全和权益风险纳入人工智能治理范围。因此，T3的共同关注主要体现为治理对象的相近性，即数据处理及相关权利风险均构成两国AI安全治理的重要政策对象。

从T5和T3可以看出，中美在风险治理领域的共同关注主要形成于治理目标和治理对象层面，为跨国风险识别和政策沟通提供了议题基础。

3.3　标准测试的技术交集与关注不对称

T1呈现出不同于T3和T5的跨国关系。该主题MIX为0.597，表明中美政策命题在统一语义空间中具有较为明显的交叉；但BAL仅为0.325，CAI为0.054，中国和美国的文档加权关注度分别为0.165和0.032，表明双方在政策权重上仍存在较大差异。从政策内容看，中国围绕人工智能标准体系、安全规范、模型与产品测试、质量评价和能力成熟度等形成了较为系统的政策安排。《国家新一代人工智能标准体系建设指

DOI：10.13998/j.cnki.issn1002-1248.26-0594

南》[28]和《国家人工智能产业综合标准化体系建设指南（2024版）》[29]均将标准、测试评估和全生命周期安全要求纳入人工智能标准体系。美国政策同样涉及技术标准、性能指标、基准测试和测量方法，*National Artificial Intelligence Research and Development Strategic Plan 2023 Update*[27]提出通过标准、指标和测试方法衡量AI系统的可靠性、安全性和可信性。

两国在标准、测试、评估和基准等治理工具上存在较明显交集，但这些工具在两国政策体系中的相对权重存在差异。中国政策较突出标准体系建设及规范化要求，美国相关内容更多体现为风险测量、性能评价和可信性验证工具。T1的跨国关联因而主要形成于治理工具层面，其较高MIX反映技术手段的相似性，较低BAL则反映这些工具在两国政策结构中的相对位置不一致。该结果表明，标准测试可以构成跨治理体系开展技术协调的政策基础，但当前证据主要支持治理工具层面的交集，尚不足以表明双方已经形成稳定的制度性合作基础。

3.4 治理重心的结构性分化

T0、T4和T2均表现出较低的BAL和CAI，但3类主题的分化具有不同政策含义。T0集中在中国政策。该主题中国文档加权关注度为0.388，为中国所有主题最高，而美国仅为0.020，BAL为0.099。中国政策主要通过内容标识、算法备案、安全评估、用户提示和服务提供者责任等方式，将AI安全要求嵌入具体服务和应用过程。《生成式人工智能服务管理暂行办法》《人工智能生成合成内容标识办法》以及算法推荐和深度合成相关政策均体现了这一治理取向。美国虽然也存在AI生成内容披露、自动化系统告知以及版权登记等要求，但在总体政策结构中的相对权重较低。T0的分化主要体现为应用环节监管工具及合规责任配置的不同侧重。

T4呈现相反结构。美国加权关注度为0.367，中国为0.021，BAL为0.108。美国政策较多涉及联邦机构AI治理能力、人才培训、政府采用、负责任研发、采购管理以及政府、产业和科研机构之间的协作，*Memorandum M*-25-21[30]等文件均体现了治理能力和人工智能应用相结合的政策安排。中国相关命题更多涉及全球治理、多边合作、治理框架和国际交流。T4的差异主要体现在政策功能配置上，美国政策赋予政府治理能力和创新应用较高权重，中国相关内容在当前样本中的集中程度相对较低。

T2在美国政策中形成独立主题，美国文档加权关注度为0.127，中国政策样本未形成对应独立主题。美国相关文本集中涉及前沿AI基础设施、数据中心、能源保障、国防和情报系统以及跨部门协调，相关行政命令和国家安全AI政策形成了较为集中的治理命题。与T0和T4相比，T2的分化具有更明显的战略安全属性，体现人工智能基础设施、前沿能力和国家安全政策之间的紧密联系。其中，中国侧未形成独立主题仅表示相关内容在当前样本中没有形成对应的集中语义集群，而不是中国不存在人工智能基础设施或国家安全相关政策。

4 中美人工智能安全治理合作的机制路径

4.1 以共同风险议题形成合作起点

具有明显跨境影响并受到双方持续关注的风险议题，是当前较具现实基础的合作领域。通过上述分析结果可知，T5具有最高的共同关注强度，T3具有最高的双边平衡程度，表明风险识别与全周期治理、数据安全与权利保护在两国政策体系中均具有较高关注度。此类议题直接涉及人工智能风险的跨境传播及共同治理需求，且与核心战略资源的关联程度相对有限，可优先纳入专业合作议程。

风险事件记录和治理证据交流可以构成基础性合作内容。*Science*一项研究指出，不良事件报告数据库能够记录现实中的具体危害，为人工智能治理积累可共享的经验性证据[31]。中美可围绕具有跨境影响的风险事件、公开评估报告和治理案例开展结构化记录和专业交流，为风险识别和比较评估提供可比较的证据基础。

风险合作还需要提高风险分类和评价框架的可比性。双方即使关注相近风险，在风险分类、严重程度及责任边界方面仍可能存在差异。风险分类、基本术语和评价维度可以作为早期协调内容，使双方能够围

绕相近风险形成可比较的问题界定和证据表达。

合作议题的优先次序还受到战略敏感程度影响。跨境外溢明显、战略资源关联程度较低的议题，可以开展较深入的专业合作；涉及前沿能力、关键基础设施和国家安全的风险，则需要控制信息交换范围和合作深度。

4.2 以标准测试与评估构建技术协调接口

T1具有较高的主题内部混合度，但两国文档加权关注存在明显差异。这说明双方在标准、测试、指标和评价方法等治理工具上存在技术性交集，但相关工具在两国政策体系中的位置并不相同。标准测试具有较明确的技术程序和较强的可比较性，可在保留各自制度安排的条件下承担技术协调功能。2026年9月发布的《人工智能安全治理框架3.0》进一步提出完善人工智能安全测评体系，相关要求为标准、测试与评估领域开展技术协调提供了政策依据[32]。

基础术语和评价维度可以作为技术协调的起点。人工智能安全测试涉及能力、可靠性、鲁棒性、滥用风险、隐私和偏差等多个维度，不同政策体系采用的术语和评价方式不完全一致。建立术语映射、风险分类对照和评价维度对应关系，有助于提高风险信息的可比性。

测试方法和结果表达也是技术协调的重要内容。测试对象界定、数据记录方式、风险报告格式和结果解释规则具有较强的技术属性，通过技术机构和专业组织开展方法比较与程序协调，可以提高不同测试体系之间结果识别和解释的一致程度。

标准协调应根据技术成熟度和政策条件分阶段推进。早期可集中于术语、指标和测试方法的兼容；在相关技术基础和互信条件较为成熟的领域，可探索评价程序和标准衔接。涉及前沿模型能力、关键基础设施或国家安全的测试内容，应根据议题敏感程度限定协调范围。

4.3 根据议题属性实施合作分层

中美人工智能安全治理议题的合作条件可以从3个方面加以判断：一是跨境外溢性，主要反映相关风险是否具有超出单一国家政策边界的共同治理需求；二是治理工具可协调性，主要考察术语、标准、测试、评价等政策工具是否具备跨制度比较和技术兼容条件；三是战略敏感程度，主要根据议题与前沿能力、关键基础设施、国防安全等领域的关联程度判断。3个维度共同影响合作的必要性、可操作性和适宜深度。结合主题比较结果，本文将相关议题归纳为基础合作层、技术协调层、审慎沟通层和制度比较层（表3）。

表3 中美人工智能安全治理议题的分层合作路径

Table 3 Tiered cooperation pathways for AI safety governance issues between China and the United States

合作层级	分层依据	对应议题	合作重点	主要执行主体
基础合作层	跨境外溢性较强，共同关注程度较高，战略敏感程度相对较低	T3、T5	风险事件记录、案例交流、风险分类、公开安全评估、数据与权利保护经验	高校科研机构、专业协会、专家网络、多边平台
技术协调层	治理工具具有较强可比性和程序化特征，但政策权重及制度安排存在差异	T1	术语映射、测试方法、评价指标、风险报告和技术标准	标准组织、测试机构、专业技术组织
审慎沟通层	与前沿能力、关键基础设施和国家安全联系紧密，战略敏感程度较高	T2及T4部分议题	风险外溢控制、基本安全边界、危机沟通、有限信息交换	政府部门、授权专家机制、多边安全对话平台
制度比较层	主要体现国内监管工具和治理体系差异，跨国规则统一需求较低	T0及T4部分议题	政策比较、治理经验交流、政策效果研究	高校科研机构、智库、二轨政策网络

基础合作层主要适用于风险治理、数据安全和权利保护等共同关注程度较高的议题。此类风险具有较明显的跨境影响，与核心战略资源的直接关联相对有限，可以开展风险研究、案例共享、公开评估和专业方法交流。高校、科研机构、专业协会及多边专业平台可以承担主要的风险研究和专业交流功能。

技术协调层聚焦具有共同治理工具的议题。标准测试虽然存在明显技术交集，但两国政策权重和制度安排仍有差异，合作重点应放在术语、指标、测试程序和评价方法的兼容。标准组织、测试机构和专业技

DOI：10.13998/j.cnki.issn1002-1248.26-0594

术网络较适合承担此类工作。

审慎沟通层主要涉及前沿人工智能能力、算力、能源、关键基础设施和国家安全等高敏感议题。相关议题战略敏感程度较高，合作目标应集中在重大风险外溢控制、意外升级防范、基本安全边界和危机沟通。具有政治授权的政府间渠道及限定范围的专业交流机制更适合承担此类沟通。

制度比较层主要涉及内容标识、平台责任、服务合规以及部分公共部门治理和创新应用议题。这些内容与国内制度结构和监管体系联系紧密，政策协调可以侧重规则比较、实施经验和治理效果研究。高校、智库和二轨网络可以通过政策比较与专业交流增进对不同治理模式的理解，无需以规则趋同作为合作目标。

4类合作层级对应不同的风险属性和制度条件。合作深度由跨境治理需求、工具协调条件和战略敏感程度共同决定，信息交换范围、规则协调强度和制度化程度也应相应调整。

4.4 构建多层次、分功能的合作机制

不同合作层级需要相应的合作机制和执行主体。中美人工智能安全治理同时涉及风险知识、技术标准、公共政策和国家安全，政府部门、科学专家网络、标准技术组织以及多边和二轨机制可以根据议题属性承担不同功能。

政府间沟通主要承担战略敏感程度较高的议题的政治协调和风险控制。涉及国家安全、关键基础设施和重大跨境风险的议题，需要具有政治授权的渠道明确基本立场、安全边界和危机处置原则。该机制与审慎沟通层相对应，其主要功能是保持最低程度的政策可预期性和危机沟通能力。

科学与专家网络主要承担基础合作层的风险研究和专业评估。高校、科研机构、专业协会及专家网络可以围绕风险事件、治理案例和评估方法开展持续研究和交流，为共同风险识别提供专业证据基础[33]。

标准与技术组织主要承担技术协调层的治理工具协调。围绕术语、测试方法、评价指标和风险报告，可发挥标准机构、测试机构及专业技术组织的专业优势，提高不同治理体系之间技术结果的可解释性和可比较性。

学术、智库及二轨网络可以承担制度比较层的政策研究和经验交流，并可在正式双边合作条件有限时维持战略敏感程度较低议题的沟通。多边机制适用于具有全球公共风险属性的议题，可通过更广泛的国际参与开展风险知识交流和规则讨论。

5 结论与展望

本文基于2020—2026年中美人工智能安全治理政策文本，采用联合主题识别与跨国比较方法，考察两国治理议题的共同关注与结构差异。研究发现，中美人工智能安全治理呈现共同风险关注、有限技术交集和政策重心分化并存的特征。AI风险识别、评估与全周期治理以及数据安全、隐私保护与权利保障具有较高的双边关注程度；标准化、测试与评估形成一定的治理工具交集；透明披露与合规责任、治理能力建设以及国家安全导向的AI基础设施则呈现较明显的政策重心差异。共同风险关注为治理合作提供了议题基础，具体合作方式仍受到治理工具和战略环境的影响。风险治理和数据权利保护可作为专业合作的优先领域，标准、测试与评估可承担技术协调功能；不同议题的合作深度可根据跨境外溢性、治理工具可协调性和战略敏感程度加以区分，并由政府间沟通、科学专家网络、标准技术组织及多边和二轨机制承担相应功能。本文主要考察中央或联邦层面的正式政策文本，研究结果反映的是政策议题结构及其潜在合作空间，尚不能直接代表政策执行效果和实际外交合作过程。未来可进一步结合双边对话、国际组织文件、专家网络及具体合作案例，考察不同议题由共同关注转化为实际治理合作的条件及其动态变化。

参考文献：

[1] 宁琪, 尹建国. 人工智能场景化应用中的风险生成及治理[J]. 改革, 2026(7): 120-130.
Ning Qi, Yin Jianguo. Risk generation and its governance in scenario-based applications of artificial intelligence[J]. Reform, 2026(7): 120-130.

[2] 祁昊天. 超越二元论: 跨国人工智能竞合关系的风险理论[J]. 国际政治研究, 2025, 46(3): 56-79.
Qi Haotian. Beyond dualism: A risk-based theory of transnational AI coopetition relations[J]. International Politics Quarterly, 2025, 46(3): 56-79.

[3] 习近平. 携手构建公正合理的全球人工智能治理体系——在2026世界人工智能大会暨人工智能全球治理高级别会议开幕式

上的主旨讲话[EB/OL]. [2026-08-26]. https://www.gov.cn/gongbao/2026/issue_12886/202607/content_7076954.html.
[4] Cai Cuihong. Shared safety responsibility[EB/OL]. [2026-08-27]. https://europe.chinadaily.com.cn/a/202605/14/WS6a05de8da310d6866eb48af3.html.
[5] The bletchley declaration by countries attending the AI safety summit, 1-2 November 2023[EB/OL]. [2026-08-26]. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.
[6] 中美举行人工智能政府间对话首次会议[EB/OL]. [2026-08-26]. https://news.cctv.com/2024/05/15/ARTIIvG55tknLsj93CHilQdX240515.shtml.
[7] 韩啸. 人工智能大国治理何以不同?——基于中美欧的元治理解释[J]. 当代亚太, 2026(3): 62-91, 186.
Han Xiao. Why do major powers' AI governance differ? - A meta-governance explanation based on China, the United States, and the European union[J]. Journal of Contemporary Asia-Pacific Studies, 2026(3): 62-91, 186.
[8] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines[J]. Nature Machine Intelligence, 2019, 1(9): 389-399.
[9] 褚松燕, 孙幼龄. 中国人工智能伦理治理的政策规制: 图景与逻辑[J]. 中国行政管理, 2026, 42(4): 35-46.
Chu Songyan, Sun Youling. Regulation of AI ethical governance in China: Landscape and logic[J]. Chinese Public Administration, 2026, 42(4): 35-46.
[10] 王少. 全球人工智能伦理治理的实现进路——以人工智能科技竞争为视角[J]. 科学管理研究, 2026, 44(1): 101-107.
Wang Shao. The pathway to global artificial intelligence ethics governance: From the perspective of artificial intelligence technology competition[J]. Scientific Management Research, 2026, 44(1): 101-107.
[11] Bengio Y, Hinton G, Yao A, et al. Managing extreme AI risks amid rapid progress[J]. Science, 2024, 384(6698): 842-845.
[12] 郭继荣, 张倩. 中美欧AI伦理政策的安全化话语对比研究[J]. 外语电化教学, 2026(2): 67-71, 86, 117.
Guo Jirong, Zhang Qian. A comparative study of securitization discourse in AI ethics policies: China, the US, and the EU[J]. Technology Enhanced Foreign Language Education, 2026(2): 67-71, 86, 117.
[13] Siddiqui S, Ibrahim L, Loke K, et al. Promising topics for US-China dialogues on AI risks and governance[C]//Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM, 2025: 1211-1223.
[14] Cihon P, Maas M M, Kemp L. Fragmentation and the future: Investigating architectures for international AI governance[J]. Global Policy, 2020, 11(5): 545-556.
[15] Tallberg J, Erman E, Furendal M, et al. The global governance of artificial intelligence: Next steps for empirical and normative research[J]. International Studies Review, 2023, 25(3): 1-23.
[16] Roberts H, Hine E, Taddeo M, et al. Global AI governance: Barriers and pathways forward[J]. International Affairs, 2024, 100(3): 1275-1286.
[17] 王净宇, 梁正. 风险类型与国际公共产品供给模式——全球人工智能安全治理的分析框架[J]. 中国社会科学, 2026(6): 94-113.
Wang Jingyu, Liang Zheng. Risk types and modes of international public goods provision - An analytical framework for global AI safety governance[J]. Social Sciences in China, 2026(6): 94-113.
[18] Wasil A R, Barnett P, Gerovitch M, et al. Governing dual-use technologies: Case studies of international security agreements and lessons for AI governance[PP/OL]. arXiv (2024-09-04)[2026-08-26]. https://doi.org/10.48550/arXiv.2409.02779.
[19] Arnold Z, Schiff D S, Schiff K J, et al. Introducing the AI governance and regulatory archive (AGORA): An analytic infrastructure for navigating the emerging AI governance landscape[C]//Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM, 2025: 39-48.
[20] 国家新一代人工智能治理专业委员会.《新一代人工智能伦理规范》发布[EB/OL]. [2026-08-26]. https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html.
[21] Blueprint for an AI bill of rights[EB/OL]. [2026-08-21]. https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/.
[22] Praveen Gujjar J, Prasanna Kumar H R. Text similarity technique using BGE-M3[C]//2025 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE). Piscataway, New Jersey: IEEE, 2025: 1-7.
[23] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure[PP/OL]. arXiv (2022-03-11) [2026-08-26]. https://doi.org/10.48550/arXiv.2203.05794.
[24] Naser M Z, Naser A Z. SPINEX-clustering: Similarity-based predictions with explainable neighbors exploration for clustering problems[J]. Cluster Computing, 2025, 28(5): 335.
[25] 全国网络安全标准化技术委员会, 国家计算机网络应急技术处理协调中心.《人工智能安全治理框架》2.0版发布[EB/OL]. [2026-08-26]. https://www.cac.gov.cn/2025-09/15/c_1759653448369123.htm.
[26] 全国网络安全标准化技术委员会.《人工智能应用伦理安全指引1.0》发布[EB/OL]. [2026-08-26]. https://www.cac.gov.cn/2026-05/22/c_1781191242686496.htm.
[27] Select committee on artificial intelligence of the national science and technology council. national artificial intelligence research and development strategic plan 2023 update[EB/OL]. [2026-08-26]. https://www.nitrd.gov/pubs/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf.
[28] 国家标准化管理委员会, 中央网信办, 国家发展改革委, 等. 国家新一代人工智能标准体系建设指南[EB/OL]. [2026-08-21]. https://www.gov.cn/zhengce/zhengceku/2020-08/09/content_5533454.htm.
[29] 工业和信息化部, 中央网络安全和信息化委员会办公室, 国家发展和改革委员会. 国家人工智能产业综合标准化体系建设指南(2024 版) [EB/OL]. [2026-08-26]. https://www.gov.cn/zhengce/zhengceku/202407/content_6960720.htm.
[30] Department of state compliance plan for OMB memorandum M-25-21[EB/OL]. [2026-08-26]. https://www.state.gov/wp-content/uploads/2025/09/DOS-Compliance-Plan-with-M-25-21.pdf.
[31] Bommasani R, Arora S, Chayes J, et al. Advancing science- and evidence-based AI policy[J]. Science, 2025, 389(6759): 459-461.
[32] 全国网络安全标准化技术委员会.《人工智能安全治理框架3.0》发布[EB/OL]. [2026-09-20]. https://www.cac.gov.cn/2026-09/14/c_1791137092283345.htm.
[33] Bengio Y, Clare S, Prunkl C, et al. International AI safety report 2026[PP/OL]. arXiv (2026-02-24)[2026-08-26]. https://doi.org/10.48550/arXiv.2602.21012.

农业图书情报学报
DOI：10.13998/j.cnki.issn1002-1248.26-0594

From Shared Risks to Governance Cooperation: Policy Issue Structure and Cooperation Pathways in China-U.S. AI Safety Governance

YU Haiyan[1], CUI Wenbo[1,2*]
(1. School of Economics and Management, East China Normal University, Shanghai 200062; 2. National Education Macro Policy Research Institute, East China Normal University, Shanghai 200062)

Abstract：[Purpose/Significance] Artificial intelligence (AI) safety risks increasingly exhibit cross-border spillover effects, and the governance choices of China and the United States have important implications for global AI safety governance. This study aims to identify the policy issue structure and cross-national relationships in China-U.S. on this basis, AI safety governance and, on this basis, examine potential areas of cooperation under conditions of strategic competition. Compared with studies that mainly focus on broad governance models or individual policy differences, this study analyzes the relative position and cross-national relationship of specific AI safety governance issues within a unified policy framework and links the empirical comparison of policy issues with differentiated cooperation pathways.[Method/Process] Based on 50 AI governance policy documents issued by China and the United States between 2020 and 2026, 4 188 governance propositions were extracted through text screening and pre-processing. These propositions were treated as the basic units of analysis. To support cross-national semantic comparison, Chinese-language propositions were converted into English semantic representations, while policy texts originally written in English were retained. BGE-M3 was used to generate semantic embeddings, and BERTopic was applied for joint topic identification. To reduce the influence of differences in document length and proposition volume between the two countries, a country–document balanced topic discovery strategy was adopted, with repeated runs under different random seeds used to assess clustering consistency. The final model identified six governance topics. For cross-national comparison, document-weighted attention was calculated, together with MIX, BAL, and CAI. MIX measures the extent to which a topic contains policy content from both countries, BAL reflects the relative balance of policy attention between the two countries, and CAI captures the actual intensity of shared attention. [Results/Conclusions] The results identify six major governance topics. AI risk identification, assessment, and lifecycle governance, together with data security, privacy protection, and rights protection, receive relatively high bilateral attention. Standardization, testing, and evaluation show a certain degree of overlap in governance tools, but their policy weights remain asymmetric. By contrast, transparency and compliance responsibility, governance capacity building, and national-security-oriented AI infrastructure display more pronounced differences in policy priorities. These findings indicate that China-U.S. AI safety governance is characterized by the coexistence of shared risk concerns, limited technical overlap, and differentiated policy priorities. Based on the characteristics of different issues, cooperation may first focus on shared risks with clear cross-border effects and relatively low strategic sensitivity, including professional exchanges, risk evidence sharing, and comparative assessment. Standards, testing, and evaluation can serve as technical coordination interfaces because they offer relatively comparable governance tools and procedures. Issues closely associated with frontier capabilities, critical infrastructure, and national security are more suitable for limited communication, risk-control arrangements, and crisis-management mechanisms. The depth of cooperation should therefore be determined by cross-border spillover effects, the coordinability of governance tools, and strategic sensitivity, with government communication channels, scientific and expert networks, standards and technical organizations, multilateral mechanisms, and Track II networks assuming corresponding functions. This study is limited to formal central- and federal-level policy documents, so the findings reflect policy issue structures and potential cooperation space rather than implementation outcomes. Future research may incorporate bilateral dialogues, international organization documents, expert networks, and concrete cooperation cases to examine how shared policy attention can be translated into sustained governance cooperation.

Keywords：AI safety; China-U.S. relations; policy text analysis; governance cooperation; global AI governance