上海人工智能实验室 Shanghai Artificial Intelligence Laboratory
CIDEG
国际与公共事务学院 School of International and Public Affairs
中央广播电视总台研究院 CHINA MEDIA GROUP INSTITUTE
西岸对话 West Bund Dialogue

人工智能安全作为全球公共产品
研究报告

AI Safety as Global Public Goods
Working Report

07.**05**
中国・上海

CONTENTS

目 录

引言

人工智能可能带来极大收益，但也可能带来系统风险，而其快速演化的技术迭代速度正在加剧发展的不确定性。这一现象已经引起了全球利益相关方的普遍关注，多边、多方主体具有共识性地认为，当前应对人工智能风险的治理措施选择至关重要，是实现人工智能可信赖与负责任的关键。

为不断提升人工智能技术的可靠性、可控性、公平性，加强人工智能安全研究、推动人工智能安全合作的重要性得到了各方肯定。人工智能安全是释放人工智能变革潜力、控制人工智能系统风险的关键措施。起草人工智能安全科学报告、形成人工智能安全评测基准、加强人工智能安全研究投入、成立人工智能安全研究机构、促进人工智能安全对话等都已成为近年来人工智能国际治理的重要行动与进程。

我们认可多边、多方主体在人工智能安全方面开展行动的积极和重要意义，同时我们也认识到当前进程的可能改善空间。基于中国实践经验，我们提出，在推动人工智能安全实践、形成人工智能安全规则、促进人工智能安全共识的过程中，应将人工智能安全视为“全球公共产品”，致力于建设公共性的安全知识、安全能力、安全资源。

“人工智能安全作为全球公共产品”的理念价值是其能够统筹发展与安全的双重目标，不仅仅将安全视为风险管理措施或监管手段，而是同时强调安全知识、能力、资源的积累与增长，以促进人工智能更好发展。该概念能够兼顾国内与国际，既有利于引导各国国内人工智能安全治理机制和体系改革，同时也有利于推动人工智能安全的国际合作与全球治理。

“人工智能安全作为全球公共产品”的理念

“人工智能安全作为全球公共产品”是指，考虑到人工智能技术创新、产业应用规律及其引致风险机制的特殊性，人工智能安全应被视为具有公共属性的知识、能力、资源，政府、企业、第三方等不同主体需要在共商共建共享过程中共同探索，促进关于人工智能安全知识、能力、资源的增长和积累。这一公共产品可更好指引人工智能产业发展与社会应用，确保每个人及人类整体从中获益，形成“小河有水大河满”的良性生态。

人工智能安全作为全球公共产品的理念认为，人工智能安全在消费上应具有非竞争性、在收益上应具有非排他性：前者是指，人工智能安全知识、能力、资源并不因新增主体而降低原有主体的使用与消费；后者是指，人工智能安全知识、能力、资源在使某些主体受益的同时，不排斥其他主体受益。

在全球层面，具有非竞争性、非排他性的国际公共产品较少。但人工智能安全议题的特殊性，正在要求我们将其建设成为全球公共产品。人工智能技术创新过程的“黑箱性”、动态性，人工智能应用的广泛性、普遍性，人工智能安全风险的不确定性、争议性，都对利益相关方推动人工智能安全治理提出了新挑战、新要求。

第一，与传统技术安全风险治理重点关注信息不对称问题不同，人工智能安全风险体现出明显的“共同无知”特征，技术创新和应用主体、政府监管部门、受影响群体都缺乏人工智能安全风险知识。第二，人工智能的跨领域、跨国界应用使得人工智能安全风险呈现出强外部性特征，不同主体间相互关联、相互依存，一方安全须建立在他方安全基础之上，任何一方都不能独善其身。第三，人工智能作为人类发展新领域的变革潜力，使得发展议题与安全议题紧密关联，发展与安全并重原则要求人工智能安全治理需要被置于发展进程中动态演化、敏捷调适。

正因为人工智能安全及其治理的新特征、新要求，人工智能安全知识、能力、资源才具有了“公地喜剧”属性。像“狂欢节”一样，越多主体参与其中，节日的气氛才越浓烈，每个人从中获得的欢乐与收益也才越大。由此，人工智能安全要求多边、多方主体的开放合作，以创造、维系有利于自身、也有利于整体的公共产品。

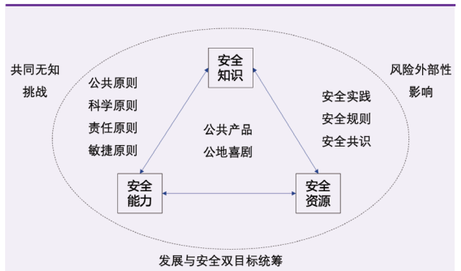


图 “人工智能安全作为全球公共产品”理念框架图

“人工智能安全作为全球公共产品”的原则

为实现“人工智能安全作为全球公共产品”的目标，应在尊重多边、多方自主决策的基础上，充分发挥各国政府、联合国及其他政府间国际组织、国际社会其他公共机构等的引导、桥接、催化作用，推动国际社会在人工智能安全领域形成能力共建、风险共担、体系共商、知识共享的可信治理环境，以克服集体行动困境，实现人工智能安全作为全球公共产品的可持续生产、维系与再生产。

实现“人工智能安全作为全球公共产品”理念的关键原则包括四个方面：公共原则、科学原则、责任原则、敏捷原则。

公共原则

“公共原则”是指，多边、多方主体要秉持全球公共性原则，在人工智能安全风险治理过程中注重开放性、包容性、普惠性，合作建设具有公信力的治理机构或跨机构的合作治理机制，共同分享、相互学习人工智能安全风险知识和治理经验，推动形成发挥公共价值作用的安全管理框架或工具，以多种形式帮助中小企业或发展中国家提升人工智能安全能力。

科学原则

“科学原则”是指，应遵循、尊重人工智能技术创新、应用的科学规律，调动企业、科学家、技术社群、监管机构、公众等利益相关方积极性，在技术创新应用过程中，共同探索、学习、积累人工智能安全知识，渐进演化、逐步提升人工智能安全共识水平，并在此过程中培养人工智能安全能力、建设人工智能安全资源。

责任原则

“责任原则”是指，要形成多类型、具有不同约束力程度的责任治理框架和制度安排，根据具体安全风险等级形成相匹配的责任制度要求，以激励相容的责任理念促进利益相关方更主动、积极参与人工智能安全公共产品建设，确保公共产品不被滥用、误用，形成良性生态体系。

敏捷原则

“敏捷原则”是指，多边、多方主体应致力于形成相互关联、相互依存的治理关系，基于宽泛共识，及时回应人工智能安全风险，以反馈互动、迭代创新的安全治理过程应对人工智能安全风险及其演化的动态性与不确定性。

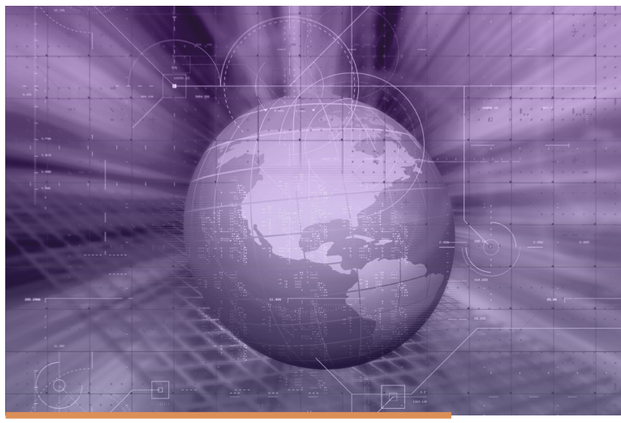
“人工智能安全作为全球公共产品”的全球价值

“人工智能安全作为全球公共产品”是基于中国人工智能安全治理实践的经验总结和理论提炼，其可作为各国国内人工智能安全治理改革、人工智能安全国际治理进程的有益补充，对在全球范围内提升人工智能安全水平具有重要的理论价值和实践指导意义。

全球范围内当前推进人工智能安全治理的制度模式和改革进程可分为三类。第一，以美国国家标准及技术研究所（National Institute of Standards and Technology, NIST）提出的人工智能风险管理框架为代表，从治理（Govern）、匹配（Match）、测度（Measure）、管理（Manage）这四个维度展开人工智能安全治理，基本特征是将人工智能安全视为风险管理的对象。第二，以英国人工智能安全研究所（UK AI Safety Institute）开发的一系列人工智能风险管理技术工具为代表，旨在通过利益相关方自主开发的技术方案来回应人工智能安全风险，基本特征是将人工智能安全视为技术风险问题加以管理、回应。第三，以欧盟《人工智能法案》中确立的风险分类管理、风险评估体系等系列制度安排为代表，基本特征是将人工智能安全视为被监管对象加以规制。

以上三种模式具有积极意义和重要价值。相比于此，“人工智能安全作为全球公共产品”概念的补充价值在于，其并不仅仅将人工智能安全视为风险管理、技术方案、规制对象，而是注意到了人工智能安全风险及其治理的新特征、新要求，并因此强调多边、多方主体在人工智能安全知识、能力、资源等方面展开同步探索的重要性。在此理念下，我们重视人工智能安全治理进程中，以创造公共产品为引导，促进多边、多方主体形成相互依赖、相互信任、共同合作的治理关系。

“人工智能安全作为全球公共产品”理念价值的释放要求治理体系和治理机制的逐渐完善。“人工智能安全作为全球公共产品”既不意味着政府规制的强制性要求，也不代表放任自流的自愿性行为，而是应遵从公共产品治理的基本逻辑，需要依赖利益相关方共同参与的制度设计与演化过程，以促进公共产品的形成和丰富。



中国实践，对话世界

“人工智能安全作为全球公共产品”的中国探索

以2017年发布的《新一代人工智能发展规划》为起点，中国近年来在人工智能发展、安全与治理等领域积极行动。针对新技术、新业态的安全治理风险与挑战，出台了一系列治理规则、建设了一批治理机构、形成了一套治理方案，为全球人工智能治理积累了中国经验。

人工智能安全是中国人工智能治理的重点关注对象和重要组成部分。基于人工智能技术创新和产业应用的科学规律，立足发展和安全并重的中国国情，中国在人工智能安全方面的治理行动可概括为以下四个方面。第一，推动利益相关方对话，就人工智能安全治理形成宽泛共识，提出了公平公正、尊重隐私、安全可控、共担责任、敏捷治理等人工智能安全治理的基本原则。第二，针对人工智能安全治理的具体问题迅速反应，在信息服务算法、深度合成、生成式人工智能的安全治理方面出台法律法规，提出了算法备案、算法分级分类、算法影响评估等系列安全治理机制。第三，调动科研机构、企业等一线力量，鼓励、要求开发人工智能安全治理技术方案，提升技术创新和应用主体安全能力，并在此过程中探索建设了人工智能安全标准、人工智能安全风险管理框架、人工智能安全风险评测基准、人工智能安全治理标杆案例等安全资源库。第四，积极参与人工智能安全全球治理进程，在《全球人工智能治理倡议》中提出推动安全治理国际合作的倡议、参加全球人工智能安全峰会、主动搭建世界人工智能大会对话平台等。

在推动人工智能安全治理进程中，中国积累了丰富的治理经验，其中重要的一条即是认为，政府、企业、第三方等利益相关主体均应将人工智能安全视为公共产品，在公共、科学、责任、敏捷原则指导下，共同推进人工智能安全知识、能力、资源的生产与积累，从而提升个体及整体层面的人工智能安全水平。在此过程中，不同主体扮演的角色有所差异：政府发挥框架作用，搭建激励相容的制度框架以为人工智能技术创新和应用主体提供探索空间，以公共服务的方式提供安全知识、能力、资源；企业及其他技术创新和应用主体发挥主体作用，主动探索人工智能安全治理方案，采取合理方式参与人工智能安全公共产品的生产与管理；第三方主体发挥催化作用，丰富治理机制、平衡利益诉求，完善人工智能安全公共产品生产与治理良性生态。该经验是在近年来应对人工智能大模型安全治理的过程中被逐渐挖掘、总结出来的，而上海市相关机构展开的探索性实践可被视为落实国家要求下的标杆案例。

1 人工智能大模型安全治理的新挑战

自2022年以来，大语言模型的突破带来了人工智能新一轮的技术创新浪潮，打开了通用人工智能技术演进的序幕，成为当前全球人工智能治理的关注焦点。但在充分拥抱大模型技术变革潜力的同时，人工智能安全风险也出现了新挑战。大模型在语言、计算、推理等多个维度接近乃至超过人的能力表现，同时凸显了虚假内容、深度伪造、舆论操纵以及就业冲击、权益侵蚀等多维度安全风险。但与传统技术安全风险不同，大模型作为新技术与新业态引致的安全风险具有以下三方面新挑战，而这也对安全治理提出了新要求。

第一，安全风险的不确定性与动态性特征更加明显。内容安全风险监管目标的模糊性以及大模型安全风险的演化特征，导致作为监管方的政府部门、作为被监管方的技术创新和应用主体、作为利益相关方的公众都难以提前预知安全风险，“共同无知”现象更为严重。这一新挑战要求多边、多方主体共同探索并积累安全知识与安全资源，普遍性地提升安全能力。

第二，发展与安全作为核心治理目标的对立、统一特征更加明显，并体现在各个层面。例如政府与企业之间作为监管与被监管方的诉求既存在重合、也存在冲突，不同政府部门之间的职责划分既存在交叠、也存在边界，不同企业之间也会以安全风险之名互相举报并走向恶性竞争。这些新特征在要求多方参与、责任共担的同时，也冲击了传统治理框架和治理行为。

第三，大模型安全风险治理机制日益复杂并呈现乱象，算法备案、红队测试、算法影响评估等现有机制存在制度定位不明确、测试标准碎片化、评估基准少公信等系列问题。这些问题导致现行治理机制不仅不能有效应对安全风险，甚至出现监管套利等乱象。

大模型安全治理的上述三点新挑战并不局限于当前，其反映的是人工智能技术迭代创新引发的一般性安全治理问题。近年来，上海市相关机构在此领域的治理探索与创新具有一定的标杆借鉴意义，同时体现了“人工智能安全作为全球公共产品”的基本理念。

中国实践，对话世界

2 上海应对大模型安全治理的实践经验

针对大模型安全治理挑战，上海相关机构主要探索了以下三方面治理创新，并在此过程中积累了治理经验。

第一，针对安全风险的不确定性和动态性特征，上海市监管部门转变传统监管思路，成为安全治理知识的学习者、汇聚者、传播者，以克服“共同无知”挑战。监管部门清楚地认识到任何一方都不可能掌握大模型安全治理所需要的全部知识，但其处于关键节点位置，既能面向广大技术创新应用一线主体，同时也具备监管权力、资源基础，因而在迅速了解安全风险点及其应对方法之后，通过与特定企业“一对一”调研交流、面向社会举办安全沙龙等方式，能够将局部知识扩散为全局知识，从而促进整个业态安全治理知识和能力的提升。

第二，针对发展与安全目标的对立、统一问题，上海市监管部门一边抓安全风险“底线共识”，一边创设跨部门协同机制，在为一线创新应用主体提供宽泛发展空间的同时，保留事中事后责任追究可能性的“惩罚性威慑”。上海市监管部门认识到安全治理不能忽视发展需要，且安全治理问题往往也需要在发展过程中才能得以解决，因而以主动探索、动态调整、共识驱动的方式设置底线范围，允许一线主体在底线以上的宽泛空间展开创新应用活动；同时，基于一线主体的自我承诺，监管部门保留算法备案、约谈预警、风险告知等治理手段，以使在出现严重后果的情况下仍然能够对企业追责。

第三，针对安全治理机制的碎片化问题，上海市监管部门充分调动公共科研、技术机构的监管支撑能力，并在此过程中广泛纳入利益相关方参与，以过程的开放性来实现安全评测数据库等安全资源库建设的公信力。同时，该安全评测数据库的制度定位不在于作为市场准入的前提，而是作为安全知识、安全能力积累与动态演化的载体。

3 “人工智能安全作为全球公共产品”的理念体现

上海实践体现了“人工智能安全作为全球公共产品”的基本理念。第一，监管机构致力于安全治理知识的学习、汇聚、传播，体现了公共原则与科学原则，其并不排斥任何主体受益于安全治理知识的增长过程，同时任何新主体的加入也不影响其他主体。第二，发展与安全目标的统筹坚持了责任原则与敏捷原则，在允许创新、鼓励创新的同时保留了责任追究的可信威慑。第三，安全资源库建设的开放性、安全评测制度定位的转变坚持了公共原则、科学原则、责任原则，其试图以激励相容的方式推动人工智能安全作为公共产品的持续增长与维系。

推动“人工智能安全作为全球公共产品”的行动倡议

上述报告对“人工智能安全作为全球公共产品”的理念、原则、价值、案例进行了解释。在此基础上，我们为未来多边、多方主体在人工智能安全领域的持续改革提出以下行动倡议，并提议形成“西岸对话：人工智能安全与治理对话网络”，以促进以下倡议的落实。

第一，以公共原则为准绳，推动 开放对话

引导人工智能安全知识、能力、资源的生产与再生产过程。应打破将人工智能安全视为竞争手段的传统思路，成立具有公共属性的人工智能安全研究机构，同时鼓励私人部门参与人工智能安全公共产品的生产过程。应转变政府治理思路，在深度融入人工智能技术产业生态的过程中，成为人工智能安全公共产品的学习者、汇聚者、传播者。在全球层面，应建设稳定对话交流机制，在国际社会创造人工智能安全公共产品。

第二，以科学原则为依托，提升 科学共识

将人工智能安全治理根植于人工智能技术与产业发展规律，承认安全治理的新挑战、新要求，并以此为基础寻找制度创新新空间。应按照人工智能安全治理需要，建设一系列人工智能安全公共产品资源库，包括安全风险评测数据库、安全风险治理案例集、安全研究公共基金等。在全球层面，应鼓励多边、多方主体加强人工智能安全治理研究，推动科学共识的逐步形成。

第三，以责任原则为边界，实现 可信承诺

在为人工智能技术与产业创新提供广阔空间的同时，保留、探索多种类型的责任机制，进而创设安全边界。应要求政府部门探索从透明、开放等程序性要求到惩罚、救济等实质性要求的全频谱责任机制创新，以解放人工智能安全治理的制度想象空间。应鼓励一线的技术创新应用主体主动探索人工智能安全治理措施，以回应责任性要求。在全球层面，应鼓励人工智能安全治理机构开展同行评议、政策学习，以丰富人工智能安全治理的工具箱与政策库。

第四，以敏捷原则为方法，促进 协同合作

在一定范围内针对人工智能安全风险实现迅速响应，并在此过程中建构多边、多方主体的相互依存治理关系。应要求多边、多方主体保持对人工智能安全风险演化的开放性和敏感性，能够在不确定与动态变化的治理环境下识别安全风险、探索治理回应。应鼓励创设人工智能安全风险的协同、交流机制，实现安全知识、能力、资源的快速传播与赋能应用。

起草机构

上海人工智能实验室治理研究中心

清华大学产业发展与环境治理研究中心

上海交通大学国际与公共事务学院

中央广播电视总台超高清视音频制播呈现国家重点实验室

支持机构

上海市经信委

上海市网信办

阿里研究院

主稿顾问

薛　澜　清华大学文科资深教授、苏世民书院院长

主稿专家

王迎春　上海人工智能实验室治理中心副主任、研究员

贾　开　上海交通大学国际与公共事务学院副教授

陈　玲　清华大学公共管理学院教授

赵　静　清华大学公共管理学院副教授

秦川申　上海交通大学国际与公共事务学院副教授

袁　媛　阿里研究院执行院长

傅宏宇　阿里研究院人工智能治理中心主任

梁兴洲　上海人工智能实验室与上海交通大学国际与公共事务学院联合培养博士生