人工智能软法治理的优化进路：由软法先行到软法与硬法协同*

曾雄① 梁正② 张辉**③
①北京科技大学文法学院
②清华大学人工智能国际治理研究院
③中国科学技术大学公共事务学院

摘 要：人工智能生成虚假内容、威胁个人隐私、加剧“算法黑箱”等，硬法是应对这些风险的传统规制工具。人工智能产生的风险具有快速弥散、动态难测的特点，而硬法缺乏灵活性、回应性以及规制对象过于僵化，无法独立胜任人工智能治理的任务。软法兼具灵活性、实验性以及多元主体合作等优势，与人工智能“善治”需求相适配。但软法存在规范性欠缺、权威性不足、约束力较弱以及执行无监督等缺陷。对此，应从创设、实施和监督等环节进行优化，包括健全支撑软法的硬法体系、提高软法质量、规范软法制定程序以及为软法提供执行保障机制等。

关键词：硬法；软法；人工智能；软法治理；硬法治理；生成式人工智能

DOI：10.16582/j.cnki.dzzw.2024.06.008

2022年底，ChatGPT“横空出世”，让人惊艳的生成式人工智能吸引了全球目光。人工智能发展迅猛，快到让业界知名人工智能专家陆奇也称“自己跟不上大模型时代的狂飙速度了”。随着虚假内容泛滥、个人隐私侵权、重要数据泄露等安全问题出现，人工智能背后的潜藏风险受到关注，各界要求对人工智能加强治理的呼声也日益高涨。对此，国内有学者提出了不同的治理思路，比如有学者提出通过顶层设计推进基础性人工智能立法。[1]也有学者提出了治理型监管范式。[2]这些学者没有对软法的功能机制进行细致探讨。实际上，以全球视野来看人工智能领域已经普遍出现软法治理的现象，比如2016年电气电子工程师学会（IEEE）发起“自动化和智能系统伦理的全球倡议”，2019年欧盟发布《可信人工智能伦理指南》（AI Ethic Guidelines），2021年联合国发布了193个会员国一致通过的《人工智能伦理问题建议书》，2023年7月21日，微软、谷歌、亚马逊等七家科技公司与美国拜登政府达成安全协议，自愿承诺管理人工智能风险等。亚利桑那州立大学2021年发布的一份关于软法治理报告的数据显示，在全球范围内确定了634个人工智能软法项目，其中90%是2017—2019年提出的。[3]由此可见，软法在人工智能治理中扮演着重要角色。究其原因，传统的硬法治理面临规则供应不足、治理及时性不够、治理有效性欠缺等问题。人工智能需要一个灵活和快速回应的治理机制，最大程度地提高应对变化中的风险的能力。[4]因此，需要从软法治理的角度予以回应。但缺乏法律强制约束力的软法该如何更好地发挥规制作用？在硬法、软法共存的局面下，如何协调两者之间的关系？基于对这些问题的思考，本文以人工智能治理领域的软法为研究对象，通过剖析人工智能风险，特别是最近兴起的生成式人工智能风险的特殊性，发现传统的“命令—控制”式的硬法规制策略面临失灵。伴随“更好的规制”（better regulation）理念的兴盛，政府规制向多元合作治理模式转型，软法发挥了重要作用，人工智能实现“善治”需要软法。综合域外人工智能软法治理的实践经验，特别是剖析我国人工智能软法治理现状，最后提出我国人工智能软法治理应坚持的基本原则以及提升软法治理效能的政策建议。

*基金项目：新一代人工智能国家科技重大专项“新一代人工智能风险防范与治理手段研究”（项目号：2023ZD0121700）。
**通讯作者
收稿日期：2023—11—28 修回日期：2023—12—27

一、现实困境：人工智能风险治理需求与硬法规制失灵之间冲突加剧

生成式人工智能快速、大规模发展，其产生的风险极具现实性，传统的硬法规制措施经常出现失灵。

（一）生成式人工智能加速公共风险的显现

ChatGPT之父萨姆·奥特曼（Sam Altman）多次强调，人类需要对人工智能风险做好准备。埃隆·马斯克（Elon Musk）甚至认为，人类可能只是引导加载程序，将引领硅基生命进化时代的来临。

1.生成式人工智能迅猛发展，虚假内容加剧社会分歧和冲突

由于机器学习的概率性质，生成式人工智能可能带来错误输出。比如ChatGPT生成的回复可能是一本正经地胡说八道，而且无法提供合理证据对其进行可信性的验证。随着聊天机器人的可及性和普及性大增，引发人们对虚假信息传播的担忧。欧洲刑警组织（Europol）发布的报告《面对真实？深度伪造的执法和挑战》（Facing reality？Law enforcement and the challenge of deepfakes）预测，至2026年，网络上90%的内容可能由人工智能创建，在未来我们会被伪造的信息所包围。[5]虚假信息的传播将加剧政治分歧，加剧数字社会的信息诚信困境，甚至煽动暴力行为。

2.训练大模型抓取海量数据，威胁个人隐私和数据安全

利用网络爬虫抓取数据是ChatGPT获取数据的重要方式，爬取过程中涉及安全数据和敏感数据。如果没有进行数据清洗，直接用于模型训练，会生成有害的内容。比如产生和传播误导性或有害信息，这对数据安全、个人隐私带来潜在危险。而且，大量用户数据为OpenAI所掌握，存在数据泄露的风险。大模型获得数据的方式还包括，在用户与其交互过程中，将输入数据（如Prompt指令内容）用为训练数据，用户在不知情的情况下，可能导入大量个人敏感信息、重要数据等。[6]

3.资源分配不均加剧产业发展不公，文化霸权滋生技术帝国主义

基础大模型训练成本高、部署困难，对工程能力要求很高。大型平台具有独一无二的数据优势、算力优势，可以强化对大模型的控制。该技术可能长期被财力雄厚的大型平台垄断，形成“马太效应”。人工智能还将制造新的国际不平等。[7]英语的统治地位具体化了文化霸权，催生了技术帝国主义，如大模型无法服务于某些语言和社区。[8]训练大模型需要耗费大量算力、能源和人力资本，科技欠发达地区需要承受因耗能带来的全球资源耗竭和环境污染等负担，却无法享受到新技术带来的福利，加剧风险处境的社会不平等。

4.模型开源加速技术扩散，先进技术演变为新型犯罪工具

全球的参数量超过百亿的大模型大约有45个，其中有15个开源，开源已经成为大模型开发的主流模式。[9]在低的使用门槛和改造门槛条件下，有人会利用开源模型从事非法活动，比如利用大模型编写诈骗短信和钓鱼邮件，甚至开发代码，生成病毒软件和勒索软件。欧洲刑警组织（Europol）发布的报告《ChatGPT：大语言模型对执法的影响》（ChatGPT—the impact of Large Language Models on Law Enforcement）提示：首先，ChatGPT可以起草高度逼真的文本，用于网络钓鱼。其次，ChatGPT能高速和大规模生成语音文本，会成为生产虚假信息的工具。最后，ChatGPT能生成不同编程语言的代码，成为制造恶意代码的工具。[10]甚至，有人在暗网上使用“毒”数据对大模型进行“有害”训练。

（二）以生成式人工智能技术为代表的智能化技术特性增加风险治理的挑战

与传统社会风险相比，人工智能产生的风险具有快速弥散、动态难测的特点，传统治理方式面临失灵。

1.风险的快速产生加剧技术治理的“步调难题”

生成式人工智能迭代升级的速度不断加快，现有的技术措施尚无法保证更高智能的系统安全可控，OpenAI主动暂停GPT5的研发。人类管控措施跟不上技术演进速度的现象，被称为“步调问题”，即监管无法跟上快速发展的技术的步伐。[11]德勤咨询公司的一份报告显示：“新技术曾经可以有两年的存在周期，但现在只在六个月内就会过时，而且变化的步伐没有放缓。”[12]监管机构的治理能力和专业知识是有限的，传统的政府治理框架无法管理包括人工智能在内的新兴技术。[13]具体

表现为：其一，监管机构立法或修法的速度太慢；其二，监管机构规制措施的效能无保证。立法或规制计划很快会过时，导致规制滞后。新兴技术具有发展速度快、应用广泛、产业链复杂的特征，消费者以更快的速度适应和接受新技术，这种加速的市场渗透也加剧了步调难题。比如，ChatGPT在推出后的几天时间里，其注册用户数超过100万，2个月活跃用户数高达1亿，是历史上增长最快的应用程序。这种高速渗透的特点让风险大规模发生后更加不可逆转。

2.风险的大规模弥散加剧组织监管的困境

风险社会是全球性的。大模型具备通用性和超强的泛化能力，特别是在开源模式下，人工智能加速向全球扩散。大模型的产业链条长，涉及环节多，覆盖行业宽，各参与主体都不同程度地加入到大模型的“调优”过程中，数据在全球不同司法管辖区流动，责任主体多元、分散，监管规则纷繁复杂。虚假内容和仇恨言论可以实现自动化大规模生成和扩散，应对这些风险超出了任何单一机构的管辖范围和治理能力。ChatGPT通过基于人类反馈的强化学习使模型的生成结果更符合人类预期，但这导致了模型的行为和偏好很大程度上反映的是标注人员的偏好，可能引入新的偏见。[14]全球的监管机构都缺乏足够的技术人员和专业知识快速、有效应对这一新兴技术，应对风险社会需要全球治理，迫切需要全球治理工具。

3.风险的动态难测增加监管问责的难度

生成式人工智能产品是动态的，其风险状况随着新数据、新用途和新的集成方式而变化，而现有规则无法对应用场景进行穷举。大模型的通用能力越来越强，覆盖任务的面越来越宽，输出的模态越来越多样。这种场景动态变化的特征，使得以具体场景判断风险的传统分析思路不再奏效。大模型具备多模态输入、输出的能力，这种跨场景的任务能力增加了依据传统的规则识别风险的难度。大模型加速实现横向、纵向的延展，参与生产的主体越来越多元和分散，传统的主体责任制度无法全面覆盖大模型技术参与者的权责，导致监管出现漏洞。比如ChatGPT，它采用基于Transformer的语言建模技术，这是一种深度学习技术。这种方法在响应生成过程中缺乏透明度，存在“黑箱”问题，这导致追踪决策路径变得复杂，从而使得传播虚假信息时追究责任的过程变得复杂。

（三）以传统硬法规制人工智能面临的困境

硬法规制是人工智能治理的重要方式，但硬法缺乏灵活性、回应性以及规制对象过于具象和僵化，无法独立胜任治理人工智能的任务。

1.硬法缺乏充分的灵活性，无法满足人工智能敏捷治理之需

虽然硬法有国家强制力作为保障，但硬法需要经过正式且漫长的立法程序和落地实施才能实现治理目标，制定法律需要耗费立法资源，程序比较漫长，还面临各种立法困难和阻碍。硬法属于刚性的强行性规范，内容具有确定性和稳定性，难以适时修订。硬法还具有自上而下和技术官僚的特征。[15]可见，仅依赖硬法已经无法满足人工智能敏捷治理之需。

2.硬法缺乏足够的回应性，无法满足人工智能协同治理之需

硬法规制主要通过政府机构，缺乏社会参与，回应性不足，民主性也不足。传统的监管流程是僵化的、官僚的、缺乏灵活性的。官僚机构往往以“家长制”方式进行干预、调控，这种命令式的硬法实施机制缺乏互动和协商，与当代主流的共治理念相逆，传统的硬法制度不适合新的、快速发展的新兴技术。人工智能治理是全球治理的一部分，也是社会共治的一部分，具有时代性、多层次性和多元性等特征，利益关系较为复杂，需要各国、各组织、各社会成员一起参与，仅依赖硬法已经无法满足人工智能协同治理之需。

3.硬法调整范围缺乏弹性，无法满足人工智能精准治理之需

硬法追求法律义务的相对确定性，规制主体、权利义务、法律责任相对具体和明确。大模型逐渐成为基础设施，未来会在各行业深入拓展应用，对各种场景的风险类型和风险高低无法预判，静态的规则治理思路无法奏效。虽然欧盟在制定《人工智能法案》（Artificial Intelligence Act）时，坚持基于风险的治理方法，将人工智能风险划分最小风险、有限风险、高风险和不可接

受风险四个风险等级，并匹配相应的监管措施。[16]但面对异军突起的生成式人工智能，这种硬法规制方式仍然面临失效的风险。

二、他山之石：域外人工智能软法治理现象与经验分析

面对人工智能带来的治理挑战，英国、美国和欧盟分别采取了基于原则的规制模式、基于管理的规制模式和基于风险的规制模式。在这三种规制模式中，软法都发挥了独特作用。

（一）英国：基于原则的规制

基于原则的规制（principles—based regulation）即不依赖详细的规则，而是通过更高层级的、笼统说明的原则或标准约束被规制主体。这种规制模式具有灵活性、敏捷性的优势，但面临执行困难的挑战。规制机构提出宽泛的原则后，必须匹配指南或标准，以便行业自我合规。在基于原则的规制模式主导下，英国政府强调在现阶段不会制定人工智能监管规则，而是发布指导意见，鼓励安全、公平和负责任地使用人工智能。英国先以自上而下的方式发布国家政策文件促进民间软法的形成，比如2023年3月发布《一项支持创新的人工智能监管方式》（A pro—innovation approach to AI regulation），其目的在于寻求建立社会共识，增加公众对人工智能前沿技术的信任，让技术企业更好地创新发展。在该政策文件中，英国政府提出了对人工智能监管的五项原则，具体包括：安全性和稳健性，透明和可解释，公平性，可问责和管理，可竞争性和救济性。在软法治理工具方面，英国政府提出要使用保证技术（assurance techniques）、自愿的指南和标准等。[17]2023年6月，英国内阁办公室在《公务员使用生成式人工智能的指引》中指出，必须意识到新技术的机会与风险，为英国公务员使用生成式人工智能制定了一般原则。

（二）美国：基于管理的规制

基于管理的规制（management—based regulation）指在计划阶段进行干预，要求被规制者完善内部管理，由被规制者自己制定计划和内部规则，以实现特定的公共目标。在基于管理的规制模式下，被规制主体获得更多的自我合规空间，但为了规避“监管俘获”或“监管断层”现象，规制机构需要匹配审计或认证标准，并对被规制主体的自我管理效果进行评估，因而需要依赖大量软法。在国家政策层面，2022年10月，美国发布《人工智能权利法案蓝图》（Blueprint for an AI Bill of Rights），提出以行业为主导，并结合具体应用场景的监管路径。[18]同时，美国国家标准和技术研究院（National Institute of Standards and Technology，NIST）发布《人工智能风险管理框架》（AI Risk Management Framework，AI RMF），提出将人工智能监管纳入风险管理框架中。[19]这是一份非强制性的指导文件，意在指导行业开发和部署人工智能时管理好风险。为了引导行业自律，美国政府在2023年7月和9月两次召集人工智能企业签署自愿承诺，内容包括制定标准、工具和测试，帮助确保人工智能系统安全、可靠和可信。此外，美国总统科技顾问委员会专门成立生成式人工智能工作组，指导安全、公平、负责任地开发和部署相关技术。

（三）欧盟：基于风险的规制

基于风险的规制是根据社会风险的影响和概率确定活动的优先顺序，其背后的理念是将不利治理结果减少到零并不是最佳策略，因为实现这种结果的成本过高，而且分散了对更严重问题的注意力或监管资源。欧盟《人工智能法案》提出“在考虑对权利和安全的影响的情况下计算风险”，并根据“人工智能系统可能产生的风险强度和范围”调整规则，“风险强度”指风险水平，即风险的大小。该法案的大部分内容涉及人工智能的风险管理，这集中体现了基于风险的规制方法。在该规制模式下，人工智能风险规制体系是一种面向不确定性的规制，适度预防原则是其核心原则，分级分类是其主要路径，风险评估、风险管理和风险沟通是主要内容，合作性、实验性、回应性、技术性、场景化是其规制特征，软法和硬法的搭配是其主要规制方法。为此，欧盟委员会先通过政策建议或伦理指南进行“软性”治理，之后再发布带有惩罚措施的法律，其监管“硬度”呈现出由弱变强的趋势。[20]比如早在2018年4月，发布《欧盟人工智能政策》（AI for Europe），2019年4月，发布《算法可问责和透明的一项治理框架》（A

Governance Framework for Algorithmic Accountability and Transparency），2019年4月，发布《可信任人工智能伦理指南》（Ethics Guidelines for Trustworthy AI）。基于这些软法，2021年4月，欧盟提出了《人工智能法案》。目前，欧盟已经正式通过了该法案。

综上所述，域外在人工智能软法治理上有一些共性的特征，包括都重视敏捷治理，并善用技术治理措施；坚持软法硬法相结合的治理方式，并适时将软法转化为硬法；既有自上而下的国家政策，也重视推动行业以自下而上的方式形成标准，国家宏观政策的指导是实现软法治理不可或缺的要素。

三、必然选择：软法属性与人工智能“善治”之需的多维契合

软法兼具灵活、实验、高效以及多元合作等优势，与人工智能治理的需求相适配，顺应了新技术治理的时代潮流。

（一）法律多元主义视角下对软法的界定

法律的概念本身是意义模糊的，正如哈特所言，“法律是什么”是长期以来困扰法学家的问题。[21]从法律的规范和约束功能视角看，法律的形式应该是多元的，法律的数量也不止一个，即一个社会群体必然同时受到不同的规范的约束。有的规范以比较正式的法律的形式呈现，有的规范则与此相反。在法律多元主义理论视角下，我们能更好地理解软法与硬法的界分。

“软法”这一概念出现于国际法领域，英文为“Soft Law”。1994年，弗朗西斯·施尼德（Francis Snyder）提出“软法是原则上没有法律约束力但有实际效果的行为规范”。软法有多种表述形式，如“合作规制”“自律规范”和“准规制”等。霍夫曼（Hoffman）等学者认为，“软法是不具有任何约束力或者约束力比传统硬法要弱的准法律性文件”。[22]国际人权、劳动、环境、金融等治理领域普遍使用软法，软法具体包括行为准则、伦理声明、专业指南、原则声明、认证计划、私人标准、公私伙伴关系或自愿计划。可以分别在国际法语境和国内法语境理解软法。在国际法语境下，软法的概念不易引起争议，也更好理解，即指那些不具有强制执行力的非条约性义务的条款或协议，比如国际组织的某些决议。在国内法语境下，软法的概念容易引发争论。从字面上看，如果软法不具有法律上的约束力，那就不应属于“法”的范畴。因此，对软法的理解应该扩张软法中“法”的概念，软法中的“法”包含法律多元意义上的社会规范，而不仅指那些由国家权力机关制定的法律。对软法的界定可以从两个维度进行：其一，制定主体的维度，软法包括非国家机关制定的对组织内部有约束力的社会规范，比如行业制定的操作指引或标准。其二，法律强制力的维度，软法与硬法之间的最根本区别在于是否有国家强制力保障实施。因此，即使国家机关制定或发布的非法律性的指导方针、政策、指令等也可以纳入软法范围，关键在于是否需要以国家强制力实施。也有学者不断拓展软法的外延，主张软法包括“法律、法规、规章中没有明确法律责任的条款”，即所谓硬法中的软法。[23]这种概念界定导致软法的定义非常宽泛，容易混淆软法与硬法之间的边界，可能动摇法律体系的稳定性，致使法的边界模糊，难以发挥软法应有的功能价值。

对此，本文认为不宜将法律中的若干原则性条款界定为软法，只要该法律文件属国家机构制定并具有强制力保障实施的就属于硬法，而无需再引入软法的概念对其进行拆分。软法具有以下特征：第一，软法的制定主体具有多样性，有国家机构、社会团队、行业自治组织、国际机构等。软法治理一般依赖自我服从、市场压力或社会舆论来实现一种新型规则的治理，其不以国家强制力为后盾。软法治理重视民主和开放的治理过程，不同主体可以对话与合作。因此，软法具有灵活、实验、多元合作等特点。以软法治理，可以创造更多的协作关系而非对立关系。第二，软法的制定程序简易、灵活、高效。软法可以由权威机构单方制定，也可以由联合体成员合意制定。软法以共识为基础，协商机制是软法制定的一个重要特征。协商代表合作、开放和自治，能够体现大部分民意，以保证软法的正当性和合法性。第三，软法本身没有严格的法定拘束力，其实施不依赖国家公权力。多数软法没有规定制裁措施，虽然少数软法规定了制裁，但这种制裁主要以民间社员身份为基

础，其最严重的是身份取缔。软法内含“应当遵守或适用”之“有效性”，是一种与“社会认同”结盟的规范性、约束性。软法无需诉诸于外部强制力量的规范，同样具备“有效性”。[24]

（二）软法的功能与效用

1.软法的功能

琳达·森登（Linda Senden）认为软法的功能包括：其一，软法可以是国家法的预备；其二，与国家法并立，相互补充；其三，解释和协助国家法。

首先，软法可以是硬法的“摆渡人”。软法向硬法转化包括法律上的转化和事实上的转化，法律上的转化包括将软法直接转化成具有法律约束力的硬法，也包括在司法程序中认可软法的法律约束力。事实上的转化指通过事实上的影响力特别是激励或惩罚确保组织成员遵守软法。[25]

其次，软法是硬法的“试金石”。与硬法的僵化不同，软法具有灵活的创设程序，能及时回应治理需求。软法可以成为硬法的“实验品”，可以不断试错和调整，给予市场更多缓冲和犯错的机会。经过试错和改进后，软法可以及时通过正式的立法程序转化为硬法。[26]

最后，软法是硬法的必要补充。监管机构制定的硬法不可能面面俱到，通常规定的是原则性内容，这样就“留下了大量的真空领域，这些领域必须或能够通过行使私性或准私性的立法权力予以填补”。[27]软法与硬法之间存在耦合关系，软法既可以是硬法的实验法，也可以是硬法的补充法。

2.软法的效用

首先，软法可以成为人工智能全球治理的工具。软法具有自愿实践和技术性较强的标准化流程，能通过领先的要素监管科技及其监管知识示范，获得其他国家的认可、学习与遵从。[28]通过技术标准或流程能以非对抗的形式达成共同遵守的治理准则，减少不同辖区的冲突与摩擦，有助于实现人工智能全球治理目标。

其次，软法降低规制成本。制定一部硬法需要走完复杂的立法程序，需要充分的调研、讨论，将耗费大量社会成本和时间成本。相较于硬法，民主协商机制使得创设软法高效，可以节约立法成本和规制成本。“软法因其制度变革的回应性、创制过程的协商性、制度安排的合意性、实施方式的温和性等特征，能够明显降低法律创制、实施与遵守的成本”。[29]

最后，规避监管技术的固化。软法治理具有阶段性，通过设定阶段性目标，反复试错、及时反馈，这样留了缓冲空间，减少了摩擦和对抗。软法治理有利于实现多元主体的参与协同，包括将各类企业、专业组织、科技团体以及社会公众等纳入人工智能治理活动中，从而避免出现单一适用硬法带来的监管技术固化问题。

（三）软法契合人工智能治理的特点与“善治”的价值追求

1999年，劳伦斯·莱斯格（Lawrence Lessig）提出“代码即法律”的论断，提出以“法律、社群规范、市场及架构”为四要素实现最佳控制、高效规制的网络空间。进入网络时代，非政府主体承担越来越多的政府职能，它们通过自我规制，干预、激励、调控其规制结果，在此过程中，规则制定、监督、执行评价集合多方参与。[30]如今，一元化的集中的垂直式权威垄断结构式微，多元化的分散的网络式权威共享结构生长，民间力量和责任感崛起，单向度的社会控制转向合作式控制。[31]按照政府主体与非政府主体在治理模式中的作用及其关系，网络空间治理可分为管理型治理、协作型治理和合作型治理这三种模式。管理型治理是传统线下监管在线上的“投射”，即监管者对被监管者的活动进行监督、管理，违反即进行规制或处罚；协作型治理是监管者的功能发挥有待网络实体的帮助和补充；合作型治理是监管者与被监管者为实现共同目标，各自发挥功能作用，参与各方的权利义务对等。进入人工智能时代，自我协商的民主管理模式进一步得到推崇。敏捷治理是实现人工智能“善治”的重要路径，灵活、韧性是主要特征。在价值目标、治理转型需求、全球治理趋势等方面，软法都与人工智能现代治理高度契合。

同时，人工智能治理具有开放性、协商性、过程性和互动性的特点，软法是实现治理的基本手段之一。首先，在开放性方面。人工智能技术是复杂的、变动的，规制主体与被规制主体同处“无知”状态，需要一起摸着石头过河，让多元的主体参与进来，消弭彼此的信息

不对称。软法的一个独特优势在于制定主体的多元化，非国家机关也能制定和实施软法。其次，在协商性和互动性方面。软法有显著的协商性和互动性特征，它往往是参与主体相互讨论和博弈的结果，体现了参与者的共同意志。最后，在过程性方面。人工智能技术演进速度非常快，出台一套规制规则可能很快就会过时，规则需要适时调整，才能适应技术发展之需。软法的制定过程简便、高效，不需要拘泥于刚性的法律程序。

人工智能技术快速迭代，开源模式成为主流，技术失控的可能性增加，其带来的风险具有复杂、多元和动态的特点。仅靠单一主体无法应对这些风险，不应再由政府单方面统揽治理资源和权力。[32]如今，“命令—控制”型的传统规制模式式微，基于风险的回应性规制（responsive risk—based regulation）[33]和精巧规制（smart regulation）兴起，这两种规制理念都强调规制手段的灵活性、参与主体的多元性，成为指导人工智能治理的重要理论。构建人工智能风险协同治理机制，通过引入多方利益相关者，与政府机构共同成为治理主体成为全球人工智能治理的重要路径。协同治理机制能“利用各自优势而逐渐组成技术治理的决策权威联盟”。[34]综上，人工智能治理领域重视发挥软法效用成为必然选择。

四、把脉问诊：我国人工智能软法治理的现状考察与问题检视

目前，我国尚未制定一部人工智能专门性法律，主要依赖一些软法实现人工智能治理之需。但这些软法普遍存在制定程序缺乏监督、内容合法性和正当性不足、难以有效执行和实施等缺陷。对此，应在软法的创设、实施和监督等阶段加以完善。

（一）我国人工智能软法的概况

我国网络立法的“四梁八柱”基本构建，包括《电子商务法》《网络安全法》《数据安全法》《个人信息保护法》等基础性、综合性、全局性法律。整体而言，我国在新技术治理方面的立法具有较强的前瞻性，但目前，我国尚未制定一部人工智能专门性法律，主要依赖一些软法实现人工智能治理之需。当前，人工智能软法主要有以下几种类型：

第一，国家层面制定的政策性文件。2017年，国务院发布《新一代人工智能发展规划》；2019年6月，发布《新一代人工智能治理原则发展负责任的人工智能》；2021年9月，发布《新一代人工智能伦理规范》；2022年，发布《关于加强科技伦理治理的意见》；2023年10月，发布《全球人工智能治理倡议》。

第二，行业标准和技术标准。2020年7月，发布《国家新一代人工智能标准体系建设指南》；2023年4月，发布《人工智能伦理治理标准化指南》。在细化数据标注规范方面，我国出台了《人工智能面向机器学习的数据标注规程》。在算法治理方面，我国出台了《信息安全技术 机器学习算法安全评估规范》《大模型训练模型技术和应用评估方法》。

第三，行业协会自律公约或合规指南。2021年，中国网络社会组织联合发布《互联网信息服务算法自律公约》和《互联网行业从业人员职业道德准则》。2023年，首都互联网协会、中国科学院信息工程研究所协调组织有关研究机构、政府智库和平台企业力量，研究编制《北京市互联网信息服务算法推荐合规指引（2023年版）》。

第四，由行业协会或企业联合发布的倡议书。2023年4月，中国支付清算协会向行业发出谨慎使用ChatGPT的倡议。2023年6月，中文在线、中国知网、中国工人出版社等26家单位发布生成式人工智能训练数据版权的倡议书。

第五，由企业发布的伦理准则、平台规范、行业倡议、合规指引等。2019年7月，北京旷视科技有限公司发布企业自身管理标准《人工智能应用准则》《正确使用人工智能产品的倡议书》和《人工智能技术合规应用指引》。2022年9月，阿里巴巴集团宣布成立科技伦理治理委员会，提出“以人为本、普惠正直、安全可靠、隐私保护、可信可控、开放共治”六大科技伦理准则。2023年5月，抖音发布人工智能生成内容的平台规范暨行业倡议。

（二）我国人工智能软法的特征与缺陷

其一，软法由国家倡导逐渐向行业自律发展，呈现出自上而下发展的趋势，国家层面发挥“指挥棒”的作

用，凸显了政府机构的宏观指导、压力传导的作用。比如国家政策提出了人工智能伦理治理要求后，旷视科技、阿里巴巴、腾讯等科技企业相继提出企业伦理治理准则，并成立相应的伦理治理组织。但是这些科技企业在伦理审查方面的具体工作进展外界并不知悉，因缺乏外界督促，行业自律的主动性尚不充足。

其二，与政府机构密切相关的行业协会或事业单位发挥了重要作用，代表或联合企业发布了多个行业倡议或合规指引，但是这些倡议却很难真正落实到企业的日常工作之中。一个重要原因在于这些行业倡议几乎与法律原则一致，缺乏可操作性。

其三，领头的企业积极推动人工智能合规治理工作，企业自律规范、平台规则成为软法的重要来源。以生成式人工智能的标识义务为例，知乎、抖音、小红书等根据《生成式人工智能服务管理暂行办法》的要求，发布大模型内容标识的规则，并研发新的标识工具，这些领头企业的行业实践将成为国家制定相关技术标准的重要来源。

其四，多数软法属于倡议性内容，没有罚则，缺乏监督，无法保证落地实施，最终实施效果无法验证。以中国互联网行业协会为例，据不完全统计，该协会发布的自律公约或公益倡议达到几十份，但是很少能从企业那里得到这些软法实际有用的正面反馈。

其五，多数软法的协商性不够，主要是单个企业自身实践经验的总结，没有经过协商讨论，互动性、参与性不足，尚不足以形成行业共识。以技术标准为例，相关标准组织在制定标准时通常依赖业界“圈子”人脉进行内部讨论，待定稿之后，再通过标准组织的网站邀请社会各界征求意见，这种制定方式导致协商性和代表性大打折扣，特别是中小企业无法充分参与其中，在市场竞争中处于劣势。

（三）我国人工智能软法缺陷的理论透视

根据软法理论，我国人工智能软法存在上述特征与缺陷的缘由有多个方面。

1.软法与硬法之间衔接不畅

软法与硬法各有侧重，作用方式有所不同，彼此之间若能建立起有机的互动衔接机制，可以更好地发挥各自优势。以技术标准为例，其作为一种重要的规范形式，若能借助法律获得一定的强制实施力，可以更好地获得推广应用。但在实践中，人工智能技术标准与法律之间的衔接机制不通畅，在相关技术标准已经颁布实施的情况下，相关法律法规援引标准的情况不充分。以生物识别技术为例，目前我国出台了多达四十多部相关国家标准，对生物识别的术语、操作规范和技术要求等构建起了规范体系。以2021年7月出台的《汽车数据安全管理若干规定（试行）》为例，其将指纹、声纹、人脸、心律等视为生物识别特征信息。根据我国关于生物特征识别标准文件的规定，生物特征识别信息包括人脸、指纹、手型轮廓、血管图像、虹膜、声纹、DNA等，而不包括心律。这样的规定未能实现法律与国家标准的对齐，扩大了可收集的生物识别特征信息范围。[35]再如《互联网信息服务算法推荐管理规定》鼓励相关行业组织加强行业自律，建立健全行业标准、行业准则和自律管理制度，但在实践中，该条款处于“空转”状态，缺乏落地的规则设计。因此，如果硬法与软法之间缺乏对应的衔接机制，容易产生规则冲突，影响硬法的可适用性。

2.软法的制定缺乏监督

软法应该满足正当性、前瞻性、科学系统性和可操作性的要求，相比于硬法，软法的制定较为随意，无统一的制定程序。软法通常以模糊的语言表达，难以贯彻和遵守，甚至大量软法重叠或不一致，逻辑性和体系性不佳。在制定程序上缺乏监督，软法可能带来不公平问题。

首先，企业之间达成自律规范会受到竞争策略、营销模式、消费偏好等影响，软法具有一定的价值偏好。比如企业之间制定某种标准的目标可能是为了防止新的竞争者进入市场。

其次，行业标准往往由在市场上占据主导地位的企业提出，无法避免私人腐败的现象。如果软法缺乏严格的立法程序、确定的解释机制和健全的实施体系，容易出现公信力缺陷。[36]

最后，由标准机构牵头制定标准时，因国内的标准组织具有“半政府”的性质，导致协商不足。在我国，大部分标准由国家行业主管部门牵头制定，遵循着自上

而下的模式，企业参与度不够。制定标准应该遵循市场化原则，以市场需求为导向，发挥行业协会和企业的作用。特别是涉及人工智能治理的标准，涉及技术细节和指标，应该由企业设定相应的指标更为科学合理。

3.软法的合法性和科学性存在不足

2023年7月6日，中国汽车工业协会联合比亚迪、特斯拉等十六家车企共同签署《汽车行业维护公平竞争市场秩序承诺书》，其中提到“规范市场营销活动，维护公平竞争秩序，不以非正常价格扰乱市场公平竞争秩序”，该条款引发公众对其违法的质疑和讨论。之后，中国汽车工业协会发布声明，将涉及价格竞争的条款从承诺书中删除，因为有违《反垄断法》精神。这个案例是行业软法的合法性和科学性需要提升的一个缩影。

软法的内容不得与宪法、法律、法规和国家政策相违背，这是软法的最基本要求，有违法律的软法肯定无效，且应受到法律的追究。在人工智能治理初期，未形成健全的软硬法渊源体系，不仅可能出现违法内容，而且不同软法之间存在相互冲突的情况，软法的合法性和正当性容易受到质疑。特别是缺乏审查和纠正机制，无法及时对违法的或与现实不符的软法进行修正，也无法对合法有效的软法进行整理，无法发挥软法协调性和有效性的功能。此外，有一些软法过于原则，形同虚设，缺乏实操标准，无法解决具体问题，甚至对国外的原则照搬照抄，未考虑中国国情。

4.软法的有效执行缺乏保障

软法最大的问题在于适用权威性不足，缺乏强制执行力。软法欠缺约束力的原因有：

其一，软法内容的不确定性消解了其自身的约束力。比如人工智能伦理准则的概念内涵都比较模糊，缺乏治理标准，很难发挥约束效力。

其二，在软法中没有明确规定具体的责任与惩罚机制。如上文所述，2021年亚利桑那州立大学发布的一份关于软法治理的调研报告发现，软法的主要特征是自愿性，同时这是软法的一个主要劣势，因为60%的软法项目没有公开列出执法或实施机制。

其三，软法不具备强制约束力，既不能成为监管机构的执法依据，也无法成为司法机构判案的裁量依据。软法理论的危险在于先验地排除了软法的司法救济。[37]由于软法缺乏法律约束力，软法的有效执行需要有健全的硬法体系、良好的信息透明机制和高效的行政协作机制作为支撑。否则，软法会因自身的妥协性，加深集体行动的困境。软法举措被当成“洗白”（whitewashing）或“洗绿”（greenwashing）的工具。人工智能伦理治理就存在“伦理洗涤”的现象。有研究表明，通过调查二十二套伦理指南后，发现人工智能伦理在很多层面是失败的，由于缺乏任何执行保障机制，伦理原则往往成为“营销”自我治理和游说的手段。[38]如2019年，谷歌成立伦理委员会，但是该小组没有实际权力否决项目。“无牙齿”的伦理原则对于科技企业而言是一个福音，因为可以以此规避监管手段或避免立法。[39]我国的情况类似，虽然旷视科技、腾讯、阿里巴巴等科技企业都建立了伦理委员会，但并未发现任何伦理审查方面的案例或透明度报告。

五、对症下药：我国人工智能软法治理的优化进路

为了充分发挥软法的效用，弥补软法治理的缺陷，应在软法的创设、实施和监督等阶段，健全软硬法渊源体系、提高软法质量、规范软法制定程序以及为软法提供执行保障机制。

（一）促进软硬法的互动衔接，健全软硬法的渊源体系

如果没有硬法作为基础性规范，软法难以发挥有效的作用。[40]

其一，条件成熟时，需要推动软法向硬法转化。

其二，及时回顾和评估软法，实现软法与硬法相互配合与相互协调，重视两者之间的衔接适用。我国需要加快人工智能立法的步伐，为人工智能软法治理提供坚实保障。需要通过法律制度建立人工智能伦理审查机制，将伦理治理工作法制化，比如尽快实施《科技伦理审查办法》，避免伦理原则成为企业躲避监管的“护身符”。

其三，建立软法统一援引机制。对于国家标准而言，立法过程中应该在术语、技术指标等方面进行援引，而不必另辟蹊径，避免规则的冲突，毕竟行业已经

按照标准进行了实践。通过援引的衔接机制，能够实现硬法与软法的贯穿融合。

（二）坚持正当程序原则，增强软法制定程序的规范性

第一，完善的立法协商机制。制定软法应该有多元主体参与，体现民主性、过程性和自愿性。[41]应最大限度地吸收利益相关群体参与讨论，表达诉求。

第二，建立立法指导机制。创设软法不能违背硬法，保证与整个法律体系兼容。

第三，建立备案与审查机制。通过备案方式，防止软法实施上的形式主义，防止软法成为逃避硬法监管的“保护伞”。还应对备案的软法建立审查机制，对违反法治精神或与硬法冲突的软法及时做修订，避免规则的冲突与对抗。比如及时将行业协会、人工智能企业和标准组织制定的软法报送至监管部门审查，对违法“软法”及时修订和废除。[42]

针对不同主体制定的软法应该形成对应的程序规范，对于政府层面制定软法，应该坚持“法无授权不可为”的总原则，明确权力清单和责任清单制度，加强立法审查，广泛听取社会意见，避免监管机构以“软法律实现硬监管”，绕开正常的立法程序，将监管之手伸得过长。对于市场主体制定软法，应提高软法的专业性和合法性，不得与国家法律要求冲突。比如针对行业协会制定的软法，需要进行公平竞争审查，以避免出现违法垄断行为。

（三）坚持法治原则，提升软法内容的正当性

制定软法应该严格遵守《立法法》的各项立法原则，避免出现违法内容。软法需要参与主体自愿遵守，制定软法应该符合社会公众的共同认知和价值观标准。

首先，在价值理念上，让软法合乎伦理道德限度。

其次，在制定主体上，扩大参与主体，发挥协商精神，避免“家长主义”。注重引入技术专家、治理专家，缓解专家与非专业的社会公众之间的认识不对称的问题。比如OpenAI宣布推出漏洞赏金计划（Bug Bounty Program），发挥技术社群的力量，挖掘ChatGPT的安全漏洞。

最后，在内容上，软法应符合理性期待。制定软法应坚持公开、民主和透明的原则，如实反馈社会公众的意见和评议。[43]软法有效性的一个重要条件是“符合一定范围内社会对值得的、更好的‘公共善’的认知与期待”，软法制定过程的协商性、沟通性是保证软法提出的“公共善”的认可程度的重要保证。

（四）充实治理工具箱，完善软法执行保障机制

软法的实施应该受到监督。[44]国外学者也提出了“软法2.0版”的概念，指软法不仅包括规则本身，还包括执行规则的过程和保障措施，即具备推进软法持续执行的工具箱。[45]软法的保障机制有：

第一，组织体系的保障。以伦理治理为例，企业应该建立伦理委员会，将伦理准则贯彻到日常经营活动中，对外应该披露伦理审查案例和年度透明报告，接受社会公众的监督。

第二，构建合规认证机制。如果要求上游供应商和下游客户遵守软法，可以通过“供应商合规声明”（SDOC）来验证，这是美国国家标准与技术研究院（NIST）认可的一种工具，使用SDOC模型让人工智能的供应商自行声明其系统符合与预期用途、性能、安全性等因素的软法要求。2021年，英国发布《有效的人工智能保证生态路线图》（The Roadmap to An Effective AI Assurance Ecosystem），致力于建立一个全球领先的人工智能保证生态，开发一系列实现人工智能安全保证的评估技术标准。[46]在国内，比如旷视科技要求技术购买方遵守其提出的伦理要求。

第三，审计或第三方机构认证。IEEE正在推进符合伦理标准的AI审计和认证计划，被称为自主和智能系统伦理认证计划（ECPAIS），该审计和认证计划能让公司在AI透明度、问责制和算法偏差等方面获得三个独立的“标记”。

第四，设计赏罚分明的奖惩机制。一是建立惩戒机制，比如声誉处罚；二是建立激励机制，设定经济激励、社会激励和道德激励。在软法合规机制中，增加对遵守规则的合规奖励，比如考虑对遵守者减轻处罚或免除处罚。

六、结论

人工智能技术动态演化所导致的风险具有高度不确定性。不可否认的是，在一段时间内，软法将在人工智能治理问题上继续发挥核心作用，应使软法尽可能有效和可信。为了克服软法的散乱性、软弱性，应逐步建立和完善硬法支撑体系。软法与硬法之间是一种耦合的关系。[47]为了实现两者的结构性耦合，应坚持软硬法协同的二元法模式。[48]在人工智能治理过程中，应充分发挥软法作为“摆渡人”的作用，让成熟的软法及时向硬法转化。

参考文献：

[1]张凌寒. 深度合成治理的逻辑更新与体系迭代——ChatGPT等生成型人工智能治理的中国路径[J]. 法律科学（西北政法大学学报），2023(03): 38–51.

[2]张欣. 生成式人工智能的算法治理挑战与治理型监管[J]. 现代法学，2023(03): 108–123.

[3]Gutierrez C I, Marchant G. A global perspective of soft law programs for the governance of AI[EB/OL]. (2022–08–07)[2023–07–22]. https://lsi.asulaw.org/softlaw/wp–content/uploads/sites/7/2022/08/final–database–report–002–compressed.pdf.

[4]Johnson W G, Bowman D M. A survey of instruments and institutions available for the global governance of artificial intelligence[J]. IEEE Technology and Society Magazine, 2021(04): 68–76.

[5]Facing reality? Law enforcement and the challenge of deepfakes[EB/OL]. (2022–04–28)[2023–07–22]. https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf?trk=public_post_comment–text.

[6]之江实验室. 生成式大模型安全与隐私白皮书[R/OL]. (2023–06–07)[2023–07–22]. https://github.com/xiaogang00/white–paper–for–large–model–security–and–privacy.

[7]贝克 U. 风险社会：新的现代性之路[M]. 张文杰，何博闻，译. 北京：译林出版社，2018: 9.

[8]Talat Z, Névéol A, Biderman S, et al. You reap what you sow: on the challenges of bias evaluation under multilingual settings[EB/OL]. (2022–05–01)[2023–07–22]. https://aclanthology.org/2022.bigscience–1.3.

[9]大模型掀热潮中国着力打造开源生态[EB/OL]. (2023–05–29)[2023–07–22]. https://tech.gmw.cn/2023–05/29/content_36594378.htm.

[10]ChatGPT–The impact of Large Language Models on Law Enforcement[EB/OL]. (2023–03–27)[2023–07–22]. https://www.europol.europa.eu/media–press/newsroom/news/criminal–use–of–chatgpt–cautionary–tale–about–large–language–models.

[11]Marchant G E. The growing gap between emerging technologies and the law[J]. The International Library of Ethics, Law and Technology, 2011(07): 19–33.

[12]Shah S, Brody R, Olson N. The regulator of tomorrow: rulemaking and enforcement in an era of exponential change[EB/OL]. (2021–01–01)[2023–07–22]. https://www2.deloitte.com/content/dam/insights/us/articles/us–regulatory–agencies–and–technology/DUP–1178_Regulator–of–tomorrow_vFINAL.pdf.

[13]Wallach W, MarchantG E. Toward the agile and comprehensive international governance of ai and robotics[J]. Proceedings of the IEEE, 2019, 107(03): 505–508.

[14]Kovacic W E, Hyman D A. Regulating Big Tech: Lessons From the FTC's Do Not Call Rule[EB/OL]. (2022–06–21)[2023–07–22]. https://scholarship.law.gwu.edu/faculty_publications/1599/.

[15]Johnson W G. Governance tools for the second quantum revolution[J]. Jurimetrics, 2019, 59(04): 487–521.

[16]曾雄，梁正，张辉. 欧盟人工智能的规制路径及其对我国的启示——以《人工智能法案》为分析对象[J]. 电子政务，2022(09): 65.

[17]A pro–innovation approach to AI regulation[EB/OL]. (2023–06–22)[2023–07–22]. https://www.gov.uk/government/publications/ai–regulation–a–pro–innovation–approach/white–paper.

[18]Blueprint for an AI Bill of Rights[EB/OL]. (2022–10–01)[2023–07–22]. https://www.whitehouse.gov/ostp/ai–bill–of–rights/applying–the–blueprint–for–an–ai–bill–of–rights/.

[19]AI Risk Management Framework[EB/OL]. (2023–01–

26)[2023-07-22]. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[20]曾雄，梁正，张辉. 欧美算法治理实践的新发展与我国算法综合治理框架的构建[J]. 电子政务，2022(07): 68.

[21]赵英男. 法律多元主义的概念困境:涵义、成因与理论影响[J]. 环球法律评论，2022(04): 146-149.

[22]Hoffman M, Rumsey M. International and Foreign Legal Research Basic Concept: A Coursebook[M]. Martinus Nijhoff Publisers, 2007: 7.

[23]姜明安. 软法的兴起与软法之治[J]. 中国法学，2006(02): 26.

[24]沈岿. 论软法的有效性与说服力[J]. 华东政法大学学报，2022(04): 101-102.

[25]漆彤. 国际金融软法的效力与发展趋势[J]. 环球法律评论，2012(02): 157.

[26]石佑启，郑崴文. 论数字平台垄断的软法治理[J]. 公共治理研究，2022(06): 9-10.

[27]博登海默 E. 法理学：法律哲学与法律方法[M]. 邓正来，译. 北京：中国政法大学出版社，1998: 423.

[28]王兰. 全球数字金融监管异化的软法治理归正[J]. 现代法学，2021(03): 117-118.

[29]方桂荣. 集体行动困境下的环境金融软法规制[J]. 现代法学，2015(04): 117.

[30]李雨峰，邓思迪. 互联网平台侵害知识产权的新治理模式——迈向一种多元治理[J]. 重庆大学学报：社会科学版，2021(02): 161-162.

[31]罗豪才. 软法与公共治理[M]. 北京：北京大学出版社，2006: 146.

[32]张铤. 人工智能嵌入社会治理的逻辑、风险与政策应对[J]. 浙江社会科学，2022(02): 74.

[33]Black J, Baldwin R. Really responsive risk-based regulation[J]. Law & Policy, 2010, 32(02): 181-213.

[34]谭九生，杨建武. 人工智能技术的伦理风险及其协同治理[J]. 中国行政管理，2019(10): 47-48.

[35]张欣. 我国人工智能技术标准的治理效能、路径反思与因应之道[J]. 中国法律评论，2021(05): 87-88.

[36]马长山. 互联网+时代“软法之治”的问题与对策[J]. 现代法学，2016(05): 53.

[37]杨海坤，张开俊. 软法国内化的演变及其存在的问题——对“软法亦法”观点的商榷[J]. 法制与社会发展，2012(06): 119.

[38]Hagendorff T. The ethics of AI ethics: An evaluation of guidelines[J]. Minds & Machines, 2020, 30(01): 99-120.

[39]Munn L. The uselessness of AI ethics[J]. AI Ethics, 2023(03): 871-872.

[40]沈岿. 数据治理与软法[J]. 财经法学，2020(01): 10-11.

[41]石佑启，郑崴文. 论数字平台反垄断的软法治理[J]. 公共治理研究，2022(06): 10.

[42]石佑启，陈可翔. 论互联网公共领域的软法治理[J]. 行政法学研究，2018(04): 58.

[43]叶泉. 全球海洋治理的软法之治与中国的战略选择[J]. 南洋问题研究，2022(04): 40.

[44]姜明安. 软法的兴起与软法之治[J]. 中国法学，2006(02): 35-36.

[45]Marchant G, Gutierrez C I. Soft law 2.0: An agile and effective governance approach for artificial intelligence[J]. Minnesota Journal of Law, Science &Technology, 2023, 24(02): 375-424.

[46]The Roadmap to An Effective AI Assurance Ecosystem[EB/OL]. (2021-12-08)[2023-07-22]. https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem.

[47]陈耿华. 联网不正当竞争行为的软法规制——兼论软法规制与硬法规制的耦合[J]. 现代财经，2016(04): 23.

[48]董正爱. 环境风险的规制进路与范式重构——基于硬法与软法的二元构造[J]. 现代法学，2023(02): 121-124.

作者简介：

曾雄（1988—），男，北京科技大学文法学院讲师，研究方向为人工智能治理。

梁正（1975—），男，清华大学人工智能国际治理研究院副院长，研究方向为人工智能治理。

张辉（1985—），男，中国科学技术大学公共事务学院特任副研究员，研究方向为人工智能治理。