推进有未来感的人工智能立法

2024年7月4日，2024世界人工智能大会暨人工智能全球治理高级别会议在上海召开，会上发布了《人工智能全球治理上海宣言》。宣言提出，要促进人工智能发展、维护人工智能安全、构建人工智能治理体系、加强社会参与和提升公众素养，以及提升生活品质与社会福祉。其中，构建人工智能治理体系的前提在于充分的制度供给。步入数字时代以来，人工智能技术在深刻改变人们生产生活方式的同时，也带来一系列不确定性风险，需要在制度层面进行系统性回应，促进人工智能技术和产业的规范健康发展。而当下，人工智能法的立法节奏如何把握？人工智能法的定位、调整对象、治理模式等应该如何确定？都是需要进一步回答的问题。为此，在2023年学界已经发布两部人工智能法建议稿的背景下，《探索与争鸣》编辑部特邀参与其中的专家撰文，以期推动人工智能立法讨论的深化，为未来我国人工智能法的正式出台，以及参与全球人工智能治理提供理论支撑。

申卫星教授指出，人工智能产业的全面发展、中央和地方先行的立法探索，以及域外人工智能立法的相关经验，构成我国人工智能立法的现实基础。要恰切把握我国人工智能立法的定位与方向，明确人工智能的概念、基本原则、风险管理制度、透明度和备案监管措施、损害和救济机制等立法过程中的重点问题。**张凌寒**教授在对学界两部人工智能法专家建议稿进行评析的基础上，指出人工智能法作为基本法的立法定位更能发挥其高位阶的引领作用。现阶段的人工智能立法工作应该采取“立总则、廓主线、留接口、适时灵活推进”的“总则式”进路。**周辉**研究员指出，人工智能法应把握促进法、安全法与程序法三重定位。在促进法层面以鼓励性、任意性规范为主，服务支持人工智能产业发展；在安全法方面，建立两级风险模式，并设立可灵活调整的人工智能领域负面清单；在程序法方面，规定人工智能的统筹协调部门，加强规则的可操作性。**苏宇**教授强调，在调整对象上，人工智能治理以及未来的人工智能法应该厘清一个基本误区，即人工智能不能被假定为生成式人工智能。人工智能的技术路线与应用方式十分庞杂，应当对其进行适当区分，采取分层多支式的治理框架。**郑志峰**教授认为，人工智能的开发者与开发活动、提供者与提供活动、使用者与使用活动、监管者与监管活动，构成人工智能法调整对象的主客坐标，同时要关注人工智能法对外的空间效力边界及其对内的与人工智能要素法和人工智能应用法的协调。**许身健**教授强调，应该重视数据在人工智能技术发展与治理中的基础性地位，数据的收集、存储、处理和使用过程内含于人工智能开发与应用的全生命周期，并由此衍生出多重法律风险。对此，应从建立完善的数据伦理体系、推动数据技术的负责任创新、迈向数据法律的整体性治理三个层次，推进全面、系统的数据治理体系建设。**张龑**教授指出，当前对于人工智能强调“填平”“纠正”，目的是为了实现人工智能与人类的“价值对齐”，但这陷入了意志论生成主义的误区。基于既有人工智能的特质，将科技伦理框架转换为有效的治理方案时，需要采取养成主义而非决断主义。**李学尧**教授通过将人工智能伦理与生物医学伦理进行对比，指出人工智能伦理的挑战主要集中在部署和应用阶段，应该在科技伦理治理框架内进行单行立法，并在伦理审查责任主体、审查启动条件、专家构成以及审查结论等方面做出基于自身特质的规范要求。

——主持人　**孙冠豪**

面向未来的中国人工智能立法：思路与重点

申卫星，清华大学法学院教授

人工智能治理正逐步从软法治理向硬法治理迈进，我国人工智能立法也已经被明确纳入立法规划，构建一部满足中国人工智能安全治理和产业发展需求，同时符合国际关于人工智能治理共识的法律，是我们现阶段的目标。现阶段，人工智能法的出场有其必要性，主要体现在三个层面：一是护航人工智能规范健康发展。人工智能全领域应用，带来的信任风险、公平风险、失控风险、社会风险、责任风险等，都需要通过制定法律制度来化解。二是保障人工智能产业创新发展。随着生成式人工智能、判别式人工智能的技术成熟和应用落地及推广，人工智能技术和产业发展中存在的制度障碍已经越来越多，需要我们通过法律制度保障人工智能产业发展。三是助力中国作为负责任的大国参与人工智能国际治理。当下，欧美等西方国家纷纷加快布局人工智能立法，通过人工智能先行立法构建其话语体系，这限制了我国人工智能技术研发的国际合作空间，也必将对我国人工智能产品的国际市场产生巨大影响。面对这一发展趋势，我国要应势而为，积极推动人工智能立法，构筑参与人工智能全球治理的前提和基础。

我国人工智能立法的可行性

第一，人工智能产业的全面发展提供了立法的社会基础。当下，以 Bert、GPT 等大规模预训练模型为基础的算法性能不断提升，人工智能内容生成（AIGC）技术在社交媒体、广播电视等领域得到广泛应用；单点算力持续提升，算力定制化、多元化成为重要发展趋势，人工智能可以在更多的终端产品中部署并为自动驾驶、远程医疗等新技术的普及提供了便利。此外，以深度学习为代表的人工智能技术普遍使用大量的标注数据而塑造了智能认知，精准营销、自动化决策技术开始在电子商务、互联网金融、公共行政等领域得到广泛应用。欧盟《人工智能法案》已经设定了构建人工智能监管沙盒的规范和标准。[①]2023 年 11 月，西班牙率先在欧盟内建立了首个监管沙盒。紧随其后，在 2024 年中关村论坛上，中国发布了

国内首个人工智能领域的监管沙盒，首批入选的企业包括中科闻歌、拓尔思和灵犀云等。随着人工智能技术不断发展，近年来相关工程落地应用呈现加速态势。

第二，中央和地方发布的一批人工智能法规、规章形成了立法的工作基础。中央和地方近年来围绕人工智能综合立法、特定技术方向立法、特定行业立法形成了一批部门规章、司法解释、地方法规，为国家立法机关开展人工智能立法积累了工作经验。在人工智能综合立法方面，2022 年《深圳经济特区人工智能产业促进条例》《上海市促进人工智能产业发展条例》先后颁布，以促进发展为主兼顾安全治理的立法思路做出了地方探索。在特定技术方向立法方面，中央网信办 2021 年 12 月发布的《互联网信息服务算法推荐管理规定》，对算法技术治理进行了深入探索，并对高风险算法采取备案检查。2022 年 11 月发布的《互联网信息服务深度合成管理规定》对利用深度学习、虚拟现实等生成合成类

① 欧盟《人工智能法案》中有关监管沙盒的规则集中在第 53 至 54a 条。此外，监管沙盒的定义规定于第 3 条第 44h 项，监管沙盒的立法目的和具体含义规定于“鉴于部分”第 71 至 72b 条。

算法制作文本、图像、音频、视频、虚拟场景等网络信息的人工智能技术治理做出探索，并前瞻性地为 ChatGPT 等生成式人工智能技术的颠覆性应用提供了安全保障规则。《生成式人工智能服务管理暂行办法》于 2023 年 8 月开始施行，中国由此成为首个为生成式大模型专门立法的国家。此外，最高人民法院针对人脸识别技术发布的相关规定，为人脸识别技术的规制提供了法律依据。[①]在特定行业立法方面，交通部、自然资源部等部门围绕自动驾驶等领域的保障性立法，为人工智能的健康发展提供了法治保障。[②]这些工作为未来的人工智能立法奠定了扎实的基础。

第三，国际组织和欧美国家的人工智能立法提供了国际经验参照。国际组织和欧美国家已经分别组成了人工智能立法工作组，并发布了系列人工智能法律（含草案）、法律预备文件。欧盟在 2022 年 9 月通过《人工智能责任指令》并确定了针对人工智能系统所致损害的适用规则。欧盟 2024 年通过的《人工智能法案》提出一种基于风险分级的监管方法。美国在 2022 年 10 月发布《人工智能权利法案》，旨在帮助指导自动化系统的设计、使用和部署，从而保护人工智能时代的公民权利。世界经合组织在 2019 年 5 月通过了《关于人工智能的建议》，世界卫生组织在 2021 年 6 月通过了《人工智能伦理与治理指南》，联合国教科文组织在 2021 年 11 月通过了《人工智能伦理问题建议书》。以上文件同时配套有支撑性的研究报告和立法方向说明，可以为我国的人工智能立法提供参照。

我国人工智能立法的定位与方向

在我国人工智能立法工作中，确定立法的重点难点问题及明确其解决思路至关重要。

第一，人工智能立法的定位应当在促进法和规制法之间找到恰当的平衡点。人工智能立法颇受关注，其中一个核心问题在于该法的定位是加强管制还是促进发展，这一定位备受社会各界关注。如果一部法律仅仅是促进法，则形同产业政策，法律的味道不浓，立法的价值难以彰显。如果立法高估人工智能的风险而设置过重的义务，可能会影响我国人工智能的快速发展，贻误在日益激烈的国际竞争中的发展契机，这也是人工智能产业界对立法最为担忧的一点。所以，未来人工智能法的立场及定位，是宏观层面要回答的首要问题。对此，笔者认为，安全与发展并不是对立关系，不发展是最大的不安全。同时，带有风险的技术创新要在安全的前提下才具有可持续性。国内外的人工智能立法和实践前沿也表明，在人工智能立法中确定安全保障措施有利于提升应用部门和消费者对于人工智能产品的信任度，一味放纵只会导致我们的人工智能产品被污名化，最终会使技术被安全要求所扼杀。人工智能产业的调研表明，技术的应用落地需要人工智能产业促进措施，需要在立法中解决制度约束问题，为科技政策提供法律支撑。

第二，人工智能立法的形式应综合考虑统一立法和分开立法的体系完整性。对于人工智能立法的形式，我国需要考虑的是制定一部综合的法律还是分领域分行业分场景单行立法？笔者对此的回答是，人工智能立法应综合考虑统一立法和分开立法的体系完整性。具体操作就是以综合为主，以单行法为补充和配套。其中的综合立法是指由全国人大常委会制定以人工智能为主题的综合性法律，为人工智能技术发展的基本原则、一般规则、管理体制、运行机制、促进发展的综合措施、法律责任等进行统一立法。此外，这样一部综合性法律难以解决所有问题，还需要根据人工智能不同应用场景和生命周期分别进行配套法规的起草，如对自动驾驶、精准医疗、政府使用人工智能决策等场景分别开展立法。这种宏观和微观视角的协调，可以满足阶段性立法和不断完善的需求，从而适应人工智能技术的不断迭代进化的特点。考虑到人工智能技术的广泛适用性和重要性，综合立法显得尤为

①《最高人民法院关于审理使用人脸识别技术处理个人信息相关民事案件适用法律若干问题的规定》（法释 [2021]15 号）。

② 例如，《智能网联汽车道路测试与示范应用管理规范（试行）》（工信部联通装 [2021]97 号）以及《深圳经济特区智能网联汽车管理条例》。

必要。同时，为了避免法律体系冲突和市场破坏，应采取中央立法先行的策略，营造鼓励科技创新的法治环境。

第三，人工智能立法的价值是打通人工智能产业发展的痛点和堵点。科技发展原则上是市场自由竞争的结果，不需要法律进行干预，不少“促进法”往往流于形式而浪费立法资源。但是，法律制度也可以通过具有强制约束力的规则促进产业发展。一是划定责任边界以增强投资预期，如明确人工智能研发者、运营者的责任边界，在人工智能领域要坚持“避风港规则”，促进产业发展；二是解决市场资源配置低效的问题，通过立法为数据供给、算力供给和应用部署建立协调机制，推动人工智能技术的发展；三是打造行业生态，通过国家立法在财政支持、人才培养、行业激励等方面提供扶持措施，从而促进专门人工智能行业生态体系的形成。

第四，人工智能法的活动空间在于人工智能立法与现行相关法律之间的关系处理。在我国人工智能立法的过程中，明确其与现行相关法律之间的关系至关重要。这不仅有助于构建协调统一的法律体系，而且能够确保法律的适用性和前瞻性。一是在促进科技发展方面，人工智能立法应充分借鉴并吸收《科学技术进步法》的主要内容。这将确保人工智能作为科技进步的重要驱动力，能够在法律框架内得到有效的促进和支持。二是明确人工智能立法与《网络安全法》《数据安全法》和《个人信息保护法》之间的关系。它们的联系在于，都涉及数据处理行为，共同构成数字经济的法律保障体系；区别在于，人工智能立法更侧重于调整算法运行规则行为，关注人工智能系统如何自动化地模拟、辅助或取代人类行为，并创造新型的人机交互关系。在这个过程中，数据的输入输出处理仍需遵循《个人信息保护法》和《数据安全法》的规定，而算法规则本身则成为人工智能立法的核心。人工智能系统的开发流程，包括模型选择、参数训练等关键环节，对系统性能至关重要，但往往超出了传统数据处理法律的调整范围。人工智能治理的着眼点是智能社会秩序，人工智能中以算法为基础的计算模型结构实际上成为了一种虚拟社会的规则。治理虚拟空间存在一些不透明、不公平、不安全的风险，其治理机制往往不是将保护个人的控制权作为目标，而是将数字社会规则或者数字社会环境的控制权或者公共秩序作为目标。所以，新的人工智能法应当特别强调公共行政监管的逻辑，对于私人权利保护也需要更多发挥行政机关的作用。简单讲，《网络安全法》《数据安全法》《个人信息保护法》构成了人工智能立法的基础性规制，但它们都缺乏对人工智能进行规范的针对性和体系完整性，人工智能立法可以说是在这三法基础上的延伸和专门化。三是明确人工智能立法与《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》之间的关系。这三个部门规章都有很强的针对性，构成了未来人工智能立法的工作基础。对于人工智能立法与这三部规章之间的关系，笔者的建议是这三部规章的重要内容都要被吸收到未来的人工智能法当中，其具体的细致规定可以作为未来人工智能法的配套法规出现。

第五，人工智能在立法节奏上如何把握。现在是不是立法的最佳时机，一直是立法的犹疑之处。人工智能产业界不乏有“让子弹继续飞”的呼声。但是，我们必须清醒地认识到，人工智能的发展带来的风险，如自我决定的削弱、高风险技术的泛化等，需要立法来确保其在安全轨道上发展。同时，打破“数据孤岛”、建立人工智能发展与数据合理使用制度等，也需要通过立法来解决。当前，中国亟需在发展与安全两个维度为人工智能提供牵引作用，避免国际竞争中的不利局面和系统性风险。鉴于欧盟和美国已经开展了综合性的人工智能立法，中国也需要有相关的立场和举措来彰显我国人工智能产品的安全可靠性，保障产品的透明度和安全性，使技术和法

律共同迈入国际市场。考虑到立法周期长，各界对于人工智能立法的方向尚未能达成一致，建议国家尽早提出人工智能法草案供社会各界充分讨论，待时机成熟适时通过我国的《人工智能法》。

我国人工智能立法的重点问题

我国人工智能立法应当采取综合立法与行业配套立法相结合的立法体系，以立法形式满足我国人工智能治理对国家基础法律制度的需求。在我国人工智能立法的进程中，确立一系列基本原则和具体制度至关重要。这一体系的构建不仅应促进人工智能的健康发展，保护社会公共利益，还应与国际标准和实践保持协调，确保我国在全球人工智能领域中的竞争力和话语权。

一是人工智能的法定概念及其调整范围。人工智能定义应涵盖从感知到决策的自动化系统，包括机器学习、逻辑和知识系统，并界定其特征。调整范围应涵盖开发者、运营者、使用者的权利与义务，以及国内外系统的地域范围，确保法律的普遍适用性和针对性。二是人工智能法的基本原则。立法应确立鼓励创新、协同治理、公众参与、安全治理四项基本原则，旨在促进技术发展、确保社会各方的参与和监督，并建立公众对技术的信任和信心。三是建立人工智能风险管理制度。借鉴欧盟《人工智能法案》的经验，未来我国《人工智能法》应建立人工智能风险评估和分类分级治理制度，科学划分应用类别和风险等级。依据人类参与程度和对人类权益、社会秩序的影响，将风险分为不可接受、高、有限和低四类，并建立相应的监管规则和评估方法。这一制度的建立有助于对不同类型的人工智能应用采取差异化的管理措施，确保风险管理和创新发展的平衡。四是完善“透明度”和备案等监管措施。基于风险分类建立透明度和备案要求，明确备案对象、内容和流程，对高风险系统建立风险管理系统，对有限风险系统要求透明度和用户知情权，对低风险系统鼓励制定行为守则。五是健全人工智能损害救济和分担机制。为了提高受害者救济能力，立法需明确侵权责任原则，并建立责任保险制度。责任主体可能包括供应商、进口商、经销商和使用者，具体应根据控制权和风险处置能力确定。高风险系统必须购买责任保险，以确保在发生损害时，受害者能够得到及时和充分的救济。现代风险社会受害者救济不能仅依赖法律责任机制，还应建立健全人工智能发展损害救济基金，以此来分散损害风险，减轻人工智能产业发展的压力，推动人工智能的创新发展。六是完善协同监管和企业行业自律机制。落实放管服监管理念，避免监管重复和真空，建立由中央网信办统筹、多部门分工的监管机制，并发挥行业组织和社会公众的监督作用。通过技术标准、风险评估等措施，加强行业组织的自治作用，鼓励在人工智能的研发和运营过程中的公众参与及监督，形成全社会共同治理的新格局。

历史上的颠覆性技术创新都带来了新的立法，以通用人工智能为代表的新技术作为全球公认的一项颠覆性技术创新，也必然形成专门立法。国务院办公厅在 2023 年度立法工作计划中首次提出“预备提请全国人大常委会审议人工智能法草案”。这一规划的法律名称表明，中国的人工智能立法选择了综合性立法路线，同时可以包含安全治理和产业发展的双重立法目标。《十四届全国人大常委会立法规划》在一类立法规划中也提出：“推进科技创新和人工智能健康发展 ..... 要求制定、修改、废止、解释相关法律，或者需要由全国人大及其常委会作出相关决定的，适时安排审议。”这一规划兼顾了人工智能立法的紧迫性和灵活性，有利于相关部门积极推进相关工作。为此，我们要有从容应对的心态和时不我待的责任感推动中国《人工智能法》的出台。

[**本文系 2024 年度清华大学自主科研文科专项“我国人工智能立法研究”项目（2024THZWYY09）的阶段性成果。**]

中国人工智能立法需凝聚“总则式”立法共识

张凌寒，中国政法大学数据法治研究院教授

自2023年6月国务院将“人工智能法”列入立法计划以来，学界关于中国人工智能立法的讨论日益广泛深入。与此同时，全球人工智能治理活动如火如荼。在此背景下，中国如何通过人工智能立法打造中国制度名片，提升国际影响力，是需要迫切研究的课题。

为了回答这个问题，中国学术界已经先后推出《人工智能示范法（专家建议稿）》《人工智能法（学者建议稿）》两项成果，前者又分为1.0、2.0两版本。两部建议稿旨在推动讨论、凝聚共识，为我国人工智能立法提供参考。在我国明确人工智能立法导向之后，如何凝聚立法共识，应当凝聚什么样的立法共识，成为当下政产学研各界亟需讨论回答的问题。本文旨在通过梳理现有的学界研究讨论成果，结合各国和地区的人工智能立法经验，提出中国人工智能立法当下应确立的制度共识，并探讨人工智能立法的定位。基于参与国际治理与推动国内治理的需要，并对立法成本予以权衡，笔者认为，我国应凝聚人工智能的“总则式”立法共识。

中国人工智能立法的制度共识

当下政产学研各界的讨论中，以善治促发展是我国人工智能立法的共同理念。对人工智能技术、服务与应用应采取审慎包容的监管而非照搬欧盟立法思路。应结合中国在国际竞争与博弈中的地位，充分考虑人工智能中的权益保护与国际合作部分。

第一，以善治促发展是人工智能立法的理念。我国人工智能立法应以促进产业发展为主要特色，立足我国人工智能产业发展实际，坚守防范安全风险的底线，力争兼具前瞻性、先进性和本土性。我国人工智能产业“领先的追赶者”的独特国际生态地位，要求我们在技术和产业的国际竞争中以发展为制度设计的主要目标，安全问题也需要通过技术发展来回应和解决。

在立法原则层面，学界发布的两部建议稿高度一致。《人工智能法(学者建议稿)》将人工智能安全这一立法目标细化拆分为

公平公正原则、透明可解释原则、安全可问责原则、正当使用原则。尤其是创新性地提出人工介入这一基本原则，以保证在人机交互中人类始终保有主体地位。《人工智能法示范法2.0（专家建议稿）》基本原则包括：治理原则、以人为本原则、安全原则、公开透明可解释原则、可问责原则、公平平等原则、绿色原则、促进发展创新原则。

促进人工智能发展是各界人工智能治理的共同理念。因此《人工智能法（学者建议稿）》中设立“促进与发展”专章，也在监督管理、责任设置等方面对此予以充分考虑，减轻了人工智能产品和服务提供者的义务，对产业创新予以一定的容错空间和责任豁免。促进人工智能发展应针对数据、算力、算法三个领域的需求，提供更大范围的高质量数据资源和算力资源支撑。《人工智能法（学者建议稿）》提出开源生态建设、产业场景培育、数字素养等条款，以多层次、全方位塑造有利于产业发展的社会氛围。《人工智能法示范法2.0（专

家建议稿）》则建议设立人工智能特区并采取授权立法机制。二者都基于同一方向提出了不同的立法建议。此外，《人工智能法（学者建议稿）》还设置了人工智能保险制度，鼓励保险市场介入，在既有的网络安全保险、第三方责任险等传统商业险种的基础上，探索适合人工智能产业的保险产品。面对用户不当使用可能导致的虚假信息泛滥等风险，还以整体性风险治理的理念提出公民数字素养的提升，从用户端预防和控制人工智能技术安全风险。

第二，开创中国本土人工智能监管制度。欧盟《人工智能法案》基于风险大小的人工智能分级监管思路，已成为多国立法参考的对象。但欧盟的这种做法并不必然适用于我国。我国应立足实际，制定符合我国需求的人工智能分级监管路径。在建立中国本土人工智能监管体系的思路上，两部建议稿不谋而合。《人工智能法（学者建议稿）》通过区分关键人工智能、一般人工智能和特殊人工智能，避免了“一刀切”的不合理监管。在分级分类监管时，由相关部门根据技术发展、行业领域、应用场景等因素进行评估并动态调整，及时更新人工智能分级分类指南，实现敏捷治理。《人工智能示范法（专家建议稿）2.0》则提出了“负面清单制度”，对负面清单所列应用设置事前许可。

关键人工智能的重点在于动态监测和安全义务，其并未采取中国社科院发布的《人工智能法示范法 2.0（专家建议稿）》中负面清单制度中的事前许可审批制度。关键人工智能没有设置事前准入门槛，仅在接收到被主管部门认定为关键人工智能的通知后，才需履行监管平台备案等义务，体现了人工智能监管的敏捷性与灵活性。关键人工智能的义务相对于一般人工智能主要包括，开发者和提供者需要建立风险披露机制和安全事件应急处置机制，以此确保关键人工智能发展的安全性。特殊领域的人工智能则是根据不同的应用场景，分别有针对性地增加安全义务。《人工智能法（学者建议稿）》希望通过这一制度设计，为产业创新减轻事前的准入负担，在确保安全需求的同时，鼓励我国企业积极参与人工智能开发和创新，促进人工智能产业健康发展。

第三，中国人工智能立法中的权益保护与国际合作。在全球范围内，随着联合国人工智能咨询机构的成立，人工智能国际治理雏形渐显，世界主要国家和经济体纷纷在监管建设、国际合作、产业发展、政策研究等多个层面加大投入，希望争取在国际人工智能治理层面的主导权。

一方面，注重权益保护彰显人本主义，是各国人工智能立法的最大公约数。《人工智能法（学者建议稿）》在制度设计上积极贯彻我国《全球人工智能治理倡议》中“以人为本”的治理理念，高度重视科技伦理与使用者权益。《人工智能法（学者建议稿）》除了在个人权益保障专章明确保护平等权、知情权、隐私及个人信息保护、拒绝权，还特别关注到数字弱势群体可能面对的数字鸿沟，明确人工智能开发者、提供者应当专门增设面向数字弱势群体权益保护的特殊功能模块。

此外，《人工智能法（学者建议稿）》还从社会整体层面为人工智能时代公众的利益提供保障。在教育领域，学者建议稿提出推进人工智能相关学科建设，开展人工智能相关教育和培训，采取多种方式培养人工智能专业人才。为提升全民数字素养，学者建议稿支持实施全民数字素养与技能提升行动，提升公民的数字获取、制作、使用、评价、交互、分享、创新、安全保障、伦理道德等素质与能力。最后，学者建议稿筹划了政府主导、社会协同、公众参与的人工智能多元治理格局。

另一方面，人工智能治理必须通过有效的国际合作才能够实现。《人工智能法（学者建议稿）》考虑到国际人工智能治理形势，设置了国际合作专章，在人工智能安全风险全球化背景下，重视人工智能治理的国际合作与交流，积极参与制定人工智能有关国际规

则、标准，推动构建开放、公正、有效的全球人工智能治理机制。人工智能全球治理不仅关系国际合作，也关乎国际博弈。我国未来人工智能立法中应明确规定，任何国家和地区对我国人工智能开发应用采取歧视性措施的，我国有权采取相应的反制措施。这一方面可以让我国采取反制措施有法可依，维护我国的合法权益；另一方面也是向国际社会明确表达相互尊重、平等互利的立场，反对利用技术垄断和单边强制措施制造发展壁垒。

中国《人工智能法》的定位

一部法律的定位是立法理念的核心问题与前提基础，无法精准把握《人工智能法》的立法定位，不仅会导致《人工智能法》难以与现有法律规范统筹协调，还会对《人工智能法》的立法结构、具体规定以及实施效果产生不良影响。在目前学界的讨论中，《人工智能法》面临着要素法、产业法、服务应用法以及基本法的选择；而无论是要素法、产业法，还是服务应用法，似乎均无法完成《网络安全法》《个人信息保护法》《数据安全法》三部法律留给《人工智能法》独特的历史使命。结合我国人工智能治理的现实困境以及未来图景，基于综合全面的考量，“人工智能时代的基本法”才应当是对《人工智能法》的准确定位描述。

（一）《人工智能法》的定位选择

谈及《人工智能法》的立法定位选择，最重要的考量因素在于该选择是否能够正确处理《人工智能法》与《网络安全法》《个人信息保护法》《数据安全法》三部法律之间的关系。

第一，从过往要素治理的视角出发，《人工智能法》可以被定位为要素法。这意味着其调整对象将关切人工智能技术、产业的诸要素，即包括底层的信息网络、算力基础设施以及数据要素、算法模型等。该立法定位，一方面可以弥补上述三部法律所遗留的对于算力、算法、模型等要素的立法空白。另一方面，随着人工智能各要素之间的纵深融合，从单一要素视角看待人工智能已经无法对人工智能进行准确描述，《人工智能法》对于人工智能各要素的有机融合，有助于搭建综合考量各要素的全面治理框架。但针对特定技术要素进行立法的模式，存在不可忽视的局限性。其一，要素法的选择意味着《人工智能法》对人工智能的技术定位在于要素的综合体，如此静态的技术性质判断，并不能应对技术高速发展带来的不确定性。其二，《数据安全法》针对数据这一核心要素，并围绕数据的收集、存储、使用、加工、传输、提供和公开等进行了全面规范，是要素法的典型代表。如果《人工智能法》同样采取该定位模式，势必会导致立法的冗余和立法资源的浪费。

第二，考虑到人工智能实现了产业的全面部署并不断促进产业升级，《人工智能法》也可以被定位为产业促进法。产业促进法定位下的《人工智能法》，既可以体现对技术产业发展的促进导向，又可以推动人工智能在生产生活中的合理应用。以发展为立法目标的《人工智能法》在内容上基本不会与上述三部以保障安全为立法目标的法律相重合。以《数据安全法》为例，其以建立健全国家数据安全管理制度和落实数据安全保护义务为主线，对数据开发利用的相关规定过于原则，无法为人工智能开发过程中涉及的数据利用问题提供具体的规则指引。[①]但产业法的立法定位可能会导致《人工智能法》过分关注人工智能技术的产业链条与商业化运营，聚焦人工智能产业的经济利益。这一方面忽视了人工智能对社会伦理、社会关系的深层影响；另一方面，也无法满足符合作为数字基础设施的基础模型的治理需求。

第三，鉴于我国人工智能治理的前期制度基础，人工智能立法还有服务应用法的定位选择。服务应用法规制的重点是人工智能在服务应用层产生的法律问题，由于位于人工智能产业链终端的服务应用层直面终端消费者并为其提供服务，该定位

① 参见董新义、梅贻哲：《“人工智能法总则”建构原则与理念——欧盟立法经验之镜鉴》，《数字法治》2024年第2期。

下的《人工智能法》可以更好地保护用户权益，针对性地解决实际问题。但也存在着在信息内容安全等问题上与此前三部法律的条款交叉重叠的可能性。因为，当前生成式人工智能防范技术尚未成熟，服务应用层面难免因此产生虚假信息生成、传播等信息内容安全问题。而信息内容安全既是传统网络信息内容治理面临的主要问题，也是人工智能尤其是生成式人工智能亟待解决的治理难题。

（二）人工智能法应是人工智能时代的基本法

在立法起草过程中，很难规定一种万能的选择立法解决方案的方法。每一种立法选择都具有其优势和风险。[①]但相较于要素法、产业法和服务应用法的定位，将《人工智能法》作为基本法的定位更能发挥其高位阶的统领及体系化的秩序作用，并能够适用人工智能技术发展的不确定性、满足中国实际的治理需求。

首先，基本法相较于要素法、产业法和服务应用法，其内容上更具统领性、一般性与综合性。基本法定位下的《人工智能法》涵盖人工智能技术发展、产业应用以及社会伦理等方面，需确立统一的基本原则和核心价值观，而不需要像要素法聚焦人工智能的不同要素，产业法具体规定产业链的各个环节，以及服务应用法局限于服务应用单一场景。可以说，基本法是要素法、产业法和服务应用法的高层次指导。此外，基本法的定位能够很好协调与上述三部法律以及相关部门规章的关系。相较于其他的立法定位选择，基本法的定位统合了网络、数据、算法、算力、应用五大基础领域的安全与发展问题，可以更好地发挥《人工智能法》高位阶的统领作用。特别是跨多个行业、多个政府机构、多个司法管辖区域和利益相关者团体的人工智能治理场景下，基本法可以更好地满足协调治理的需求，应对人工智能所引起风险的复杂性、系统性和不可预见性。

其次，作为人工智能时代的基本法，《人工智能法》需要提供一个体系化的治理框架。体系不是法律在形式上的追求，而是法律应当所具备的“德性”之一。[②]只有坚持基本法的立法定位，才能协调人工智能治理相关法律规范和标准，保障人工智能治理体系的可预期性。基本法的定位对《人工智能法》的立法进路提出了较高的要求，需要立法者全面考虑人工智能技术的快速发展和广泛应用所带来的多重影响，但相较于要素法、产业法和服务应用法，基本法的定位能够全面、系统地考虑人工智能技术的发展现状和未来趋势，为人工智能的安全与发展提供坚实的法律基础和保障。

最后，《人工智能法》的基本法定位在确立总体框架和原则的同时，为未来的法律调整和补充预留了空间和接口。随着技术的发展和社会的变化，《人工智能法》可以灵活调整和更新，保持适应性和前瞻性。人工智能立法面临着新技术迭代升级和产业快速增长的形势，基本法的立法定位并不对人工智能技术进行价值预设，而是将人工智能视为政治、经济和技术三位一体的复合型对象。相比之下，要素法隐含着人工智能是技术要素组成的定位，没有考虑到人工智能推动数字社会生产变革的切面；产业法则是局限在数字社会生产场景下将人工智能视为推动数字经济发展的关键工具，强调“物质刺激的力量而忽视良知的力量”，[③]会导致社会问题的恶化；服务应用法的定位则是将人工智能视为一种服务与应用的模式。然而，面对人工智能所引发的技术链式突破，人工智能立法需从之前网络立法关注的服务应用层面拓展至技术研发层面与要素治理层面。[④]

中国人工智能立法的“总则式”进路

中国人工智能立法的“总则式”进路既可以有效确认中国人工智能立法现阶段各界的共识，又可以满足当下人工智能国际治理和打造中国制度名片的迫切需要，同时为未来人工智能全面立法留足空间。

① 参见海伦·赞塔基：《立法起草：规制规则的艺术与技术》，姜孝贤译，北京：法律出版社，2022年，第58页。

② 雷磊：《适于法治的法律体系模式》，《法学研究》2015年第5期。

③ 琳恩·斯托特：《培育良知：良法如何造就好人》，李心白译，北京：商务印书馆，2015年，第219页。

④ 参见张凌寒：《中国需要一部怎样的〈人工智能法〉？——中国人工智能立法的基本逻辑与制度架构》，《法律科学（西北政法大学学报）》2024年第3期。

第一，中国人工智能立法的“总则式”进路，能够有效在现阶段凝聚各方共识，确认人工智能治理的基本制度。虽然《人工智能法》应当作为人工智能时代的基本法，但立法时不应抱持“毕其功于一役”的理念，一蹴而就地出台一部大而全的法典。社会各界对如何治理人工智能技术、要素和应用等问题尚存在争议，其中不乏人工智能立法会制约技术发展的观点。因此，如何促进社会各界达成人工智能立法的共识，既解决人工智能治理的紧迫问题，又为未来的发展留足空间，都是人工智能立法需要考量的问题。基于此，现阶段《人工智能法》的立法工作应当坚持“总则式”进路，采取“立总则、廓主线、留接口、适时灵活推进”的进路。参考《民法典》的“总则式”的立法体例，可以通过“提取公因式”的方式对立法目的，人工智能技术、产品与服务必须遵循的基本原则，基本法律制度应遵循的一般性规则，以及我国人工智能治理的基本理念与基本立场进行明确。

第二，人工智能立法的“总则式”进路能够满足当下人工智能治理和打造中国制度名片的迫切需要。《人工智能法》作为中国人工智能治理体系的提纲挈领之作，应当确立人工智能治理的总则，起到价值统领作用。人工智能法调整对象是治理人工智能技术、产业和应用全链条及在此基础上形成的复杂系统与社会生态。这意味着《人工智能法》是一项系统性工程，涉及多个层面和维度。要保障人工智能立法的科学性，应当充分借鉴我国立法的成功经验。具体而言，其一，立法目的方面，应当包括促进人工智能技术创新与应用、维护社会公平与正义、推动人工智能产业健康发展等。其二，应当明确贯穿整个人工智能治理体系的基本原则。其三，在基本理念和立场方面，应当充分表达中国始终将人民的利益和福祉放在首位，积极参与国际人工智能治理合作的人本主义立场。

第三，人工智能立法的“总则式”进路为未来人工智能全面立法留足空间，也可搁置当下无法达成共识的争议。“总则式”的《人工智能法》功能在于明确治理方向和主要问题，廓清人工智能治理的主线。一方面，人工智能立法需从之前网络立法关注的应用层面拓展至要素聚合的治理层面。从技术发展的角度来看，人工智能应用随着技术的链式突破不断改变。以应用为调整对象的立法模式，难以适应技术发展给调整对象带来的不确定性。《人工智能法》可充分借鉴既往法治实践中累积的应对不确定性的技术发展、社会关系变化风险的经验，兼顾人工智能产业发展所涉的诸要素，并提炼规范人工智能要素的共性规则。首先，在算力层面，《人工智能法》应当对算力基础设施建设、算力资源利用、绿色算力发展等进一步细化。算力作为人工智能发展的基础，其建设和管理直接影响人工智能技术的应用和推广。因此，相关制度应对算力资源的公平分配、节能环保等方面作出明确规定，确保算力资源的高效利用和可持续发展。其次，在算法层面，《人工智能法》应当对透明度要求、风险评估义务、内容标识、算法备案、人工智能审计等规则予以完善。最后，在数据层面，作为人工智能的“燃料”的数据，其质量和安全性直接影响到人工智能的表现。数据投喂带来价值偏见、隐私泄露、数据污染等新的数据安全挑战，同样需要《人工智能法》的关注与回应。待到未来社会全面步入人工智能时代，监管方积累了充分的治理经验时，则可以考虑制定人工智能的综合性立法及至最终的人工智能法典。

面对人工智能的迅猛发展及其给传统法律制度带来的挑战，有效的人工智能治理刻不容缓。“良”性治理离不开“良”法施行，高位阶的人工智能立法成为各国人工智能治理的布局重点，并逐步成为各国抢抓人工智能国际治理主导权的重要发力点。基于此，本文在分析中国人工智能立法迫切性与必要性的基础上，梳理现阶段达成的立法共识与人工智能治理的迫切需求。当下，中国人工智能立法应凝聚“总则式”立法共识。

中国人工智能立法：示范与定位

周辉，中国社会科学院文化法制研究中心研究员

人工智能快速迭代、广泛渗透，既会带来伦理失序、技术失控等新风险，也会加剧技术黑箱、算法歧视等旧问题，亟待从法律上系统作出回应。中国学界先后推出的《人工智能示范法（专家建议稿）》《人工智能法（学者建议稿）》两项

成果，在为国家人工智能立法工作提供具体条文和制度设计参考的同时，也持续凝聚着政产学研各界关于促进和规范人工智能发展的法治共识。但是，学术的争鸣不能替代制度的供给。有必要加快深入分析总结人工智能发展和安全的法治需求，更好研究制定符合我国国情、顺应世界发展潮流的人工智能法律制度体系。

《人工智能示范法（专家建议稿）》的起草逻辑

起草《人工智能示范法（专家建议稿）》，既希望通过以示范法形式对未来我国的人工智能立法提供示范参考，也希望以此为基础，探讨中国人工智能治理的理论命题和范式，形成体现中国网络与信息法治特点，符合中国新兴领域发展需要的人工智能法治话语体系，也为人工智能产业的发展提供示范的合规指引。综合考量我国人工智能治理实践和人工智能发展需求，《人工智能示范法（专家建议稿）》中采取了设立人工智能专门主管机关、建立人工智能负面清单的治理模式，同时尽量避免过度超前创设新兴权利，为未来的人工智能应用划出底线、留足空间。[①]

第一，设立人工智能专门主管机关。不同于《人工智能法（学者建议稿）》在统筹协调机制下由各部门在职责范围内负责人工智能监督管理工作的思路，《人工智能示范法（专家建议稿）》提出设立国家人工智能办公室作为国家人工智能专门主管机关，同时在省级人民政府和较大的市的人民政府中设立负责辖区内人工智能治理的人工智能主管机关，将人工智能领域的规则制定、监管执法、促进产业发展等职能集中于人工智能主管机关，避免监管领域交叉重叠可能导致的“九龙治水”难题。

实践中，具体的应用场景、商业模式往往难以明确区分，某几种人工智能服务或应用可能落入多个部门的监管领域中。此时，既有可能出现各部门均进行立法、重复监管的问题，也有可能出现各部门均怠于立法、监管空白的情况。设立“统筹协调”机制解决这一问题，必然要明确统筹协调机制与各参与部门之间的权限划分，事实上仍然是由统筹协调机制作为一个整体负责部门牵头制定人工智能领域的基础性规则和通用标准，其他部门则负责结合本行业内具体情况进行执法；而相比于多个部门共同负责或者指定某个部门作为统筹协调的负责部门，设立集中统一的人工智能主管机关并由国家人工智能主管机关领导各级地方人工智能主管机关履职，可以更加系统地掌握人工智能的发展和治理情况，制定更加协调统一的人工智能规则、标准，也便于人工智能研发者、提供者与主管部门进行沟通协作。除此之外，人工智能主管机关还可以作为我国与其他国家、国际组织就人工智能治理合作进行对等交流的机构，配合外交部门提升我国参与人工智能全球治理的专业程度。当然，设置集中

① 参见周辉、李延枫：《〈人工智能示范法〉释义》，北京：中国社会科学出版社，2024年，第1—10页。

统一的人工智能主管机关，并不排斥在制定人工智能治理细则、实施监管执法时，吸收对应行业、领域的主管部门参与。

第二，以负面清单替代分类分级管理制度。不同人工智能技术的底层逻辑千差万别，如何防范、如何控制、如何消除人工智能在不同环节、不同应用场景下的风险，尚无可通用的标准答案，而只能予以场景化、类型化的类案分析。为了回应类似需求，欧盟《人工智能法案》尝试基于发生风险的概率和严重程度对人工智能进行全面的分类分级，并根据分类分级的结果禁止实时生物识别等少数人工智能应用，同时对高风险的人工智能施加较为严格的合规义务。①

但是，如果将风险评估、划分的义务更多地赋予人工智能研发者、提供者，例如要求其自行评估、判定其所研发或提供的人工智能系统的风险，则有可能迫使研发者、提供者为确保合规而笼统地按尽可能高的等级来确定人工智能系统的安全风险，以避免出现不可预料的高风险及承担相应的法律责任。此外，设置多个不同等级的风险虽然划分更加细致，但在实践中对企业合规的参考价值相对较小。为保证一定的冗余度，企业会在难以判别自身所需合规举措时主动采用更高乃至最高的风险防范标准。这种“就高不就低”的导向固然可以防患于未然，但也会损害技术创新的积极性。

因此，相较于已有的分类分级监管模式，《人工智能示范法（专家建议稿）》仅将风险明确为“高风险”“低风险”两级，同时有针对性地明确何种特定人工智能服务、应用属于高风险。在明确高风险人工智能的基础上，可由监管部门建立人工智能“负面清单”，对清单内的人工智能研发、提供等活动设置较高的准入门槛及合规义务。清单之外的人工智能研发、提供活动则仅需承担一般性的安全义务，以减轻研发者、提供者的风险防范负担。“负面清单”不需要人工智能研发者、提供者自主确定所研发和提供的人工智能应属何种范畴，未在负面清单中列举的人工智能服务和应用均视为低风险。同时，为了对可能出现的风险实施必要的防范，“负面清单”也应不断更新，及时纳入属于高风险范畴的人工智能。此种模式可以为人工智能产业提供更高的确定性，减少人工智能新技术新应用分类分级不明确而无法形成一致标准的情况，也与我国《促进和规范数据跨境流动规定》在数据跨境领域实施的“负面清单”相仿，能够在划定安全底线的前提下鼓励企业大胆探索、积极创新。

第三，审慎设置新型权利。近年来，不乏有从权利视角介入人工智能或其他新兴技术治理，主张赋予个体以若干项新型权利的研究和讨论，但从目前技术发展阶段来看，各种新型权利的内涵尚不明晰、权利实现机制尚不成熟；从学术研究角度固然可以进一步探讨其合理性，但在制度设计的维度，则应保持审慎，仅作必要的规定，并留出未来进一步修改、完善和进行制度衔接的空间。

例如，《人工智能示范法（专家建议稿）》设置了人工智能使用者要求提供者作出解释的权利，但对此种“解释权”设置了一定限制，不仅将其适用条件限于人工智能产品、服务对个人权益有重大影响的情形，还允许提供者综合考虑产品、服务的场景、性质和行业技术发展水平等因素作出反馈，而没有就解释的内容、解释方式方法等进行明确规定。这是因为，类似的权利内容在个人信息保护等现有制度中已有体现，通过现有的知情权、监督举报权等权利行使足以满足需要。当然，在人工智能发展过程中，未来可能会遇到需要在特定领域设立某项权利，或广泛赋予个人以某项权利的情况；但在技术发展尚不充分、尚不完善的情况下，不应急于通过创设更复杂的权利义务关系解决发展中尚不明晰的问题。如果所设立的权利在实践中无法得到有效实现，或与现有权利内容重复度较高，则既无益于保障权利，也无益于促进发展。

① *Artificial Intelligence Act*, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf, 2024 年 7 月 1 日访问。

《人工智能法》的三重定位

结合现有的研究与实践，未来的《人工智能法》应主要把握促进法、安全法与程序法三重定位。

（一）鼓励发展创新的促进法

与侧重于禁止性、命令性规定的管理型立法不同，促进型立法在设范方式上更加灵活，以鼓励性、任意性规范为主，从服务而非监管的角度，引导、支持特定领域或事业的发展。

第一，在数据供应方面，要在扩大数据供给的同时，降低人工智能训练过程中数据合理使用的制度门槛。人类生产生活过程中产生的数据看似无穷无尽，但经筛选与加工以后真正对人工智能训练有实际价值的数据却并非如此。图书、文献等高质量数据的更新速度远小于网络空间内的信息流规模，在人工智能训练需求不断扩张的趋势下，高质量数据到 2027 年就有用尽的风险，甚至低质量数据也可能在本世纪中叶消耗殆尽。[①]因此，既要在现有共享数据库、公用数据库基础上进一步支持人工智能领域基础数据库和专题数据库，扩大用于人工智能训练的公共数据供给范围，也要明确人工智能训练数据合理使用制度、个人信息合理使用制度，减轻人工智能研发者承担的不必要的合规义务。例如，经去标识化技术处理的个人信息在理论上仍可借助额外信息复原，因此《个人信息保护法》仅将这一技术作为安全措施的一种，个人信息处理者即使采用了去标识化技术也不能免于承担《个人信息保护法》规定的其他义务。但是，在人工智能训练场景中，触达训练用原始数据的权限较为可控，个人信息因与其他数据混同用于训练而复原难度大大增加，实际上已几乎不能从人工智能输出的结果反向识别出其训练过程可能涉及的特定个人，个人信息权益受影响风险较其他场景显著降低。故可考虑在人工智能立法中针对人工智能训练出台专门规定，明确人工智能研发者、提供者采用符合国家标准的去标识化技术对个人信息进行处理并仅用于人工智能训练时，可豁免其履行个人信息处理者义务。

第二，在算力供给方面，应通过建立公共算力资源供给制度、推动公共算力资源平台建设与利用、加强算力科学调度等方式，解决企业尤其是中小型企业在人工智能研发过程中所面临的算力资源不足的问题。统筹算力资源既包括我国北京、上海、广州等地正在建设或已经建设的公共算力资源平台，也包括一些科研机构、大型企业为开发自有人工智能系统而筹集的算力资源。在互惠互利、公平合理的前提下，可以探索算力资源市场化交易，促进公共算力资源、私有算力资源的共享互通，提高算力资源利用效率。

第三，在支持算法创新方面，除了鼓励闭源模型的发展外，还要重视开源人工智能生态对人工智能技术创新的驱动作用。开源人工智能以开放式的知识共享机制为核心，不仅能够吸引中小规模研发者参与协作，也有助于大模型的优化完善。可以在立法中明确国家对建设、运营开源开发平台、开源社区的支持与鼓励，也可以探索设立开源人工智能基金，提供专项资金促进开源人工智能生态的繁荣，并可鼓励政府机关先行先试，应用符合要求的开源人工智能。

第四，出于激发人工智能创新活力、促进先行先试和控制负外部性的影响，还可以参考经济特区、浦东新区和海南自由贸易港的授权立法模式，在已有国家新一代人工智能试验区基础上作出更进一步的制度安排，选择人工智能发展要素聚集度好、立法能力强、治理水平高的“试验区”设立“人工智能特区”，建立授权立法机制。人工智能特区所在城市人民代表大会及其常务委员会，可以结合特区内人工智能创新发展实践需要，遵循宪法以及法律和行政法规的基本原则，就人工智能研发、提供、使用活动制定法规，在人工智能特区范围内实施。人工智能特区法规应当分别报全国人民代表大会常务委员会和国务院备案；对法律、行政法规的规定作变通规定的，应当说明变通

① Villalobos, Pablo, et al., *Will We Run Out of Data? An Analysis of The Limits of Scaling Datasets in Machine Learning*, http://arxiv.org/abs/2211.04325, 2024 年 7 月 1 日访问。

的情况和理由。人工智能特区法规还可以就同一事项，与部门规章或特区所在省、自治区、直辖市的地方性法规、地方政府规章作出不同规定，并优先适用。

（二）防控安全风险的安全法

突出人工智能立法的安全法定位，主要是划定人工智能研发、提供过程中的安全底线，建立人工智能安全评估、审计和应急处置制度，科学区分人工智能全生命周期中各主体的安全义务。例如，就人工智能技术自身的安全风险，在设置负面清单对高风险人工智能实施更严格监管的基础上，还可以参考《个人信息保护法》对大型平台设置制定平台内规则、处置平台内违规行为等特殊义务的做法，要求基础模型研发者对自身研发模型的安全性负更高的注意义务，并应专门采取合规措施保障基础模型安全。此外，为落实一般主体的安全防范要求，也可以根据主体从事活动的不同，区分人工智能研发者、提供者，有针对性地设置不同义务。人工智能研发者应更加注重所研发的人工智能系统在技术上的安全风险，而人工智能提供者应侧重于及时向研发者和监管部门反馈人工智能运行过程中出现的问题，关注对使用者可能造成的影响。

当然，人工智能立法需要防范的安全风险不止技术自身的潜在风险，也包括人工智能技术发展应用过程中，外部因素干扰所带来的有害信息倒灌、产业链供应链受影响等安全风险。对此，人工智能立法应与《反外国制裁法》等法律法规形成制度衔接，设计对其他国家和地区不合理措施予以必要反制的法律依据。

（三）规范权力行使的控权法

人工智能是新兴技术领域，无论是促进其发展还是对其风险进行治理都无太多先例可循，这一过程中可能产生行政权力过分干预行业发展的情况。未来在人工智能专门立法的框架下，应由人工智能主管部门及时制定类似《网信部门行政执法程序规定》的履职制度，规范与人工智能有关的行政执法程序，避免过于频繁的检查、执法和监测对人工智能研发者、提供者的经营活动造成阻碍。

同时，也需要注重人工智能治理规则的可操作性。例如，关于监管部门制定人工智能负面清单的标准以及基础模型的概念等问题，除了评估其可能对社会秩序、法秩序造成的冲击外，也需要遵循特定的技术指标。在基于清单进一步斟酌清单内外监管制度时，应避免给主管机关、人工智能研发者和提供者增加不必要的负担。对负面清单内的人工智能研发、提供活动，综合防范风险的需求，宜采取许可申请制，实行事前监管；对负面清单外的提供活动，则可以实施以备案为主的事后监管，人工智能研发活动在未转化为实际应用前则可不作备案要求。区分许可与备案、事前监管与事后监管，虽然将清单外人工智能研发与提供活动纳入监管范围，但此处所指“备案”及其他监管执法、风险监测活动，不应成为人工智能研发者、提供者正常经营的阻碍。就备案而言，对于应备案而未履行备案义务、通过不正当手段取得备案等情形，相关主体仍应承担相应的不利后果；但是，只要材料符合形式要求、信息齐备，人工智能主管机关就不应利用备案流程变相实施实质性审查，而应准予备案。在未超出法定期限的前提下，只要申报材料已提交，备案流程的完成也不应成为开展人工智能提供活动的必要前提。

总而言之，人工智能立法不仅要对新技术、新应用带来的风险挑战予以回应，更应着眼于为智能经济、智能社会构建具有基础性、原则性的制度框架。对人工智能带来的新问题新需求，应在充分研究论证的基础上作出必要规定，并留出未来修改、完善的空间；对不便在较宏观的高位阶立法中直接进行规定的细节性问题，应明确主管机关出台配套制度的期限，做好制度衔接。

［本文系中国社会科学院2024年度实验室孵化专项项目“人工智能安全治理研究”（2024SYFH007）的阶段性成果。］

人工智能的多层分支式治理框架

苏宇，中国人民公安大学数据法学研究院院长、教授

人工智能的技术路线与应用方式极为庞杂。如何对人工智能实行适当的区分式治理，是人工智能立法中最为基础和关键的问题。法律需要对人工智能作出合适的区分，避免以单一、僵化的规则调整丰富多样的人工智能对象。在国内

近期有关人工智能法治的学术讨论中，人工智能法的调整对象往往被有意或无意地假定为生成式人工智能乃至大模型，而实际上这只是整个人工智能的一个局部，远难覆盖人工智能技术与应用之全貌，以此为模板设计人工智能法律体系，极易出现“挂一漏万”或“削足适履”的后果。对此，应当根据人工智能不同技术路线的特点，推行多层级分支式治理架构，构筑我国人工智能法律治理的基本框架。

人工智能的“四层二分”框架

根据现行有效国家标准《信息技术人工智能术语》(GB/T 41867-2022，以下简称《术语》)的定义，人工智能系统是指“针对人类定义的给定目标，产生诸如内容、预测、推荐或决策等输出的一类工程系统”，而人工智能（在相关学科领域内）是指“人工智能系统相关机制和应用的研究和开发”。在人工智能的所有其他相关标准及已有行业实践中，人工智能都以这样一种广泛的含义被理解和运用。从人工智能发展史观之，人工智能最初是基于规则的人工智能（Rule-based AI）或“规则型”人工智能。此后，经过以神经网络为主要代表的联结主义的探索与实践，[①] 基于机器学习的人工智能（Machine Learning-based AI）或“学习型”人工智能崭露头角。根据《术语》的定义，“机器学习”是指“通过计算技术优化模型参数的过程，使模型的行为反映数据或经验”。机器学习向算法模型中引入了可训练的参数，也使得代码本身并不能以确定的规则揭示模型的决策逻辑，从而形成了所谓的“算法黑箱”，带来了棘手的治理难题。在采取机器学习路线的人工智能中，一部分模型只能在有限选择范围内作出十分明确的一类判断或决策结果，如人脸识别、形状检测；而另一部分模型则可根据输入的信息或指定的条件从十分宽广的选择范围内选择和组合元素以生成有意义的新信息，即生成式人工智能。在生成式人工智能的发展历程中，词嵌入技术的成熟和 Transformer 架构的运用，使语义处理技术实现了飞跃。大语言模型的诞生堪称人工智能发展史上里程碑式的革命，同时跨越不同领域、面向公众开放的通用大模型也对人工智能的法律治理带来了最大挑战。

由此，人工智能的“四层二分”框架已清晰可见。首先，根据是否包含可学习和迭代的参数，人工智能可被分为学习型人工智能和规则型人工智能。其次，根据模型是否有能力生成事先未完全指定的信息，学习型人工智能可被分为生成式人工智能和非生成式人工智能。再次，根据模型生成信息的过程中是否对包含语义的载体（尤其是文本）进行编码、解码操作，生成式人工智能可被分为处理语义的人工智能和语义无涉的人工智能。最后，根据模型是否以一定强度学习了公开途径可得数据以外的知识及信息，处理语义的人工智能可分为公用模型和特殊用途模型。每一层的划分都以确定的技术特征为基础，也都具有相应的

① 参见顾险峰：《人工智能的历史回顾和发展现状》，《自然杂志》2016 年第 3 期。

法律治理需求，可以导向治理规则的分层渐进设计，进而形塑人工智能法律体系之纲目。

各层分支的治理需求与制度设计

上述人工智能的“四层二分”框架中，风险发生逻辑、利益链条和治理思路最为简明的是规则型人工智能。在此基础上，每增加一项新的关键差异，就需要从制度层面形成相应的“治理增量”，回应相应层面的特殊治理需求，从而使我国人工智能立法形成切合人工智能技术与应用特性的多层级分支式治理框架。

（一）第一层分支：规则型人工智能与学习型人工智能

规则型人工智能通常采用决策树、线性回归、朴素贝叶斯分类器等“白箱型”算法，本质上是一个自动化决策系统。因此，规则型人工智能的治理规则可以参考对自动化决策系统的治理规则。民商事法律活动中，对（白箱型）自动化决策系统的应用本质上并不需要引入特别的法律规则，仅在涉及个人信息处理等领域时有所谓的拒绝自动化决策权等例外。涉及行政职能和公共服务领域，自动化决策系统的规制框架已有较多理论和实践探索，以“技术性正当程序”[①]为代表的相关思路可以被用于构建自动化决策系统法治化的理论与制度框架。

学习型人工智能较之规制型人工智能增加了两项重要特征：一是“算法黑箱”的出现。如果模型采取堆叠单元连续函数的方式逼近目标函数，从而实现“万有逼近”[②]（universal approximation）之能力，算法模型的黑箱型特征会更加明显。与此相应的法律治理需求就是算法解释制度群，即算法解释、算法可解释性、算法透明等制度。二是可训练参数的影响。在程序代码之外，可训练参数（权重＋偏置）对结果的影响举足轻重。鉴于人工智能模型在作出决策时整个参数张量的状态与训练数据和训练过程密切相关，训练数据来源复杂，而训练后的参数张量是否隐含偏见、歧视或其他违反法治价值的数理结构，无法被简单直观衡量。由此新增的法律治理需求聚焦以下两点：其一，训练数据和训练过程合规，国外早已出现的建设“公平数据分析与分类系统”之类的主张即典型要求。[③]我国人工智能治理方面的有关行政规章已关注此方面的制度建设，如《生成式人工智能服务管理暂行办法》（以下简称《暂行办法》）第 7 条即专条规定了训练数据处理的基本要求。其二，算法审计。算法审计不限于代码审计，尽管算法审计可以“着眼于全链条、全周期的治理”，[④]但就目前全球范围内的主要实践来看，这一机制主要被用于发现经训练的算法模型所隐含的偏见或歧视。对于部分易受价值观影响的重要算法模型，应建立算法审计制度以防止其偏离我国法律规范所认可的价值观念。

（二）第二层分支：非生成式人工智能和生成式人工智能

算法解释制度群、算法审计、训练数据与训练过程合规等已可基本满足非生成式人工智能的治理需求，生成式人工智能新增的治理需求主要涉及网络信息生态内容治理，触发《暂行办法》和《互联网信息深度合成服务管理规定》（以下简称《深度合成规定》）的监管。生成式人工智能由于可以“生产”网络信息，进入网信部门的监管范围，由此而增加的治理需求及制度回应也已体现在网信部门制定的多部行政规章中。

《暂行办法》对生成式人工智能初步建立了较为完整的治理框架。其中，“服务规范”第 9 条要求生成式人工智能服务提供者“依法承担网络信息内容生产者责任，履行网络信息安全义务”，而“网络信息内容生产者责任”来源于国家网信办 2019 年制定的《网络信息内容生态治理规定》。生成式人工智能的应用同时也触发《深度合成规定》的规制。《暂行办法》第 22 条第（一）项将生成式人工智能技术定义为“具有文本、图片、音频、视频等内容生成能力的模型及相关技术”，而《深度合成规定》第 23 条第一款将深度合成技术界定为“利用深度学习、

① 苏宇：《数字时代的技术性正当程序：理论检视与制度构建》，《法学研究》2023 年第 1 期。

② 苏宇：《算法解释制度的体系化构建》，《东方法学》2024 年第 1 期。

③ Joshua Kroll, Joanna Huey, et al., “Accountable Algorithms,” *University of Pennsylvania Law Review*, vol.165, no.3, 2017.

④ 张欣、宋雨鑫：《算法审计的制度逻辑和本土化构建》，《郑州大学学报》（哲学社会科学版）2022 年第 6 期。

虚拟现实等生成合成类算法制作文本、图像、音频、视频、虚拟场景等网络信息的技术”，并以“包括但不限于”的方式列举了六类技术。因此，这一定义在应用上实际已覆盖生成式人工智能技术，一旦生成“网络信息”，即落入《深度合成规定》的调整范围。《暂行办法》亦认可《深度合成规定》对生成式人工智能的覆盖，如其第 12 条规定了“提供者应当按照《互联网信息服务深度合成管理规定》对图片、视频等生成内容进行标识”。因此，生成式人工智能因其“制作”网络信息的能力而触发一系列监管措施，但究其实质，其中不少治理增量实际上应由第三层分支承担。

（三）第三层分支：处理语义的生成式模型和语义无涉的生成式模型

生成式人工智能包含一个内容庞杂的谱系，基于 Transformer 算法的大型语言模型只是生成式人工智能技术晚近发展出的技术路线。生成式人工智能早年的代表性算法如生成对抗网络、循环神经网络等，并不必然拥有处理自然语言文本的能力。如果一项生成式人工智能技术被用于生成无法对应现实世界的图像、纯音乐旋律、自然环境中的声音等，其引起的法律风险有限，甚至理论上并不需要触发整个针对生成式人工智能或深度合成的规制框架。然而，如果某项生成式人工智能技术有能力处理人类语言所表述的意义（典型技术如词嵌入）并使之包含于某种形式的输出中，则其可能导致的法律风险将显著提升。其不仅需要完整地接受本质上源于网络信息内容生态治理的现行多项规定的监管，还需要进一步建立和完善以测评为中心的系列治理机制，形成真正意义上的生成式人工智能生态治理体系。

在生成式人工智能中，语义无涉的生成式模型尽管可能生成违法或有害信息（如包含淫秽或暴力因素的图片），但却很少真正触及复杂的政治和社会问题。处理语义的生成式模型可以实现文本向量化（vectorization），主流的方法将文本正确地分割成标记（token），并嵌入到模型可以使用的密集浮点数表示中。[①]这使得模型有可能学习文本中不同词符及词符组合之间的关系，进而塑造模型对语义的某种数值化“认知”。只有基于一定形式的文本向量化，模型才有可能学习和处理不同形式的命题，从而形成模型自身的某种“思维链”和“价值观”，而不仅仅是非生成式模型中存在的各种局限于特定分类的“歧视”或“偏见”。如果模型能够基于丰富多样的语义针对性地调整输出，在相当程度上依赖统计学方法的算法审计就可能力有不逮；而面对规模庞大的大模型，算法解释制度群的实施面临尚未完全解决技术可行性的问题。在模型应用范围和影响力达到一定水平的条件下，需要对模型进行系统性的测评，全面检测和评价模型在价值观、安全性和关键能力等方面的表现与缺陷，这是本层分支中最关键的增量治理需求。

在此基础上，目前我国呈现高度交叠的深度合成和生成式人工智能治理规则也可以被区分开来，实现“各司其职”。深度合成服务的法律治理机制足以承载语义无涉的生成式人工智能模型的治理需求，而具备语义处理能力的生成式人工智能模型则需要专门性的生成式人工智能治理法律体系，以全面应对模型在思维链和价值观等方面的复杂风险。当然，鉴于此类模型强大的综合能力和战略价值，支持和促进发展性质的引导措施必不可少；对于规模庞大、影响广泛、支撑业务众多的关键基础模型，还需要考虑以“新型数字基础设施”的定位为其提供制度保障。

（四）第四层分支：公用模型和特殊用途模型

当前，我国学界对具备语义处理能力的生成式人工智能已给予高度关注，但主要聚焦于对不特定免费或付费用户开放的公用模型，对于特殊用途模型缺乏关注。特殊用途模型可被广泛应用于国防、外交、司法、工业、公共安全、社会管理、科学研究等领域，其主要法律风险在于：第一，特殊用途模型可能通过专门训练数据、知识推理组件、基于模型的知识编辑等方式吸纳不开放乃至保密的知识和信息，不仅数据安全风险较公用模型

① Elie Bursztein, Marina Zhang, et al., “RETVec: Resilient and Effcient Text Vectorizer,” *In the 37th International Conference on Neural Information Processing Systems*, no. 2661, 2023, p. 60902.

突出，各种评估、认证、测评、审计等第三方治理机制的开展亦备受限制。第二，特殊用途模型往往有某些性能指标的刚性约束，与公用模型的训练目标不尽相同。例如，司法大模型对回答准确率有严格的要求，但不追求与公用模型同等的有效回应率；在其选择作出实质性回答的情况下，必须杜绝编造法条、捏造规则、混淆关键概念等源于“模型幻觉”的错误。受制于大模型的技术原理，证明偏差（attestation bias）与语料库频率偏差(relative frequency bias)等幻觉来源难以完全避免，此种要求在公用模型中难以实现，[①]但在增加了控制组件的领域模型中却属可行。这就需要针对特殊用途模型建立专门性的治理机制。

特殊用途模型的治理机制主要应当是精心设计的隔离式治理机制，即保证在隔离敏感数据和专业知识的基础上，仍然尽可能实现模型的可靠与安全，以及对特定性能指标的满足。隔离式治理机制需要仔细厘定第三方专业机构（涉及国防、外交等敏感领域时甚至包括监管部门）可以介入的边界，通过模拟数据运行条件下的模型测评、算法影响评估和交付用户的参数与梯度显示功能、随机性程度调节功能、定制量化解释功能等共同实现风险治理与算法安全目标。如向模型引入的知识需要法律为其可靠性、准确性与时效性提供特殊保障，还需要建立针对张量空间的专门性监管规则，基于特殊用途模型是通过提示工程、参数编辑抑或控制组件等方法注入知识，要求采取相应的技术标准或达到特定的性能指标。

人工智能多层分支式治理的法治价值

上述“四层二分”框架旨在建构我国人工智能法律治理的基础性制度架构。认知框架的主要作用是化约复杂性，此种多层分支式框架不仅契合人工智能的发展历程和技术原理，也最大限度地减少了不同分类维度交叠引起的认知混乱和治理措施组合失配。逐层加入的“治理增量”追求风险增量与治理增量的阶段性适配，有助于提升法律治理不同技术路线人工智能应用的规制精度及合比例性。

诚然，人工智能的技术和应用无疑还存在若干具备法治价值的分类，如人工智能模型总体上包括开源模型和闭源模型，生成式人工智能模型中也包括通用模型和专用模型，公用模型中还需要划分出关键模型和普通模型。其中，尤受瞩目的是关键基础模型的判别标准。当前欧美部分制度实践以训练时消耗的浮点数计算量或运行时需要的算力为判别标准，但我国盲目移植这一标准可能会产生“刻舟求剑”的后果，因为随着人工智能技术的发展，实现与当前大模型同等能力的模型所需算力可能继续上升还是转而下降尚未可知。人工智能立法必须将“技术飞速发展给调整对象带来的不确定性”[②]作为重要考量因素。因此，尽管对于具有广泛用途和影响力的关键基础模型有作为新型数字基础设施加以特别保护的必要，也可能存在对此类可能引致重大风险的模型施加特殊规制的需要，但在相关判别标准和治理机制尚充满不确定性的情况下，此层划分尚需更多理论与实践探索，以求形成更完备的规则体系。

人工智能法治面临的难题，在于强烈的不确定性和高度复杂的生态制约了规则精度的上限，而在规则精度有限的前提下，框架结构上的重叠、缺漏与错位很难通过特殊规则的精准设计填补。“四层二分”框架不仅尽可能避免了各种复合型多维框架结构的缺陷，也有利于未来继续划分必要的层次和分支，还便于未来形成的各种人工智能治理机制顺利嵌入人工智能法制，避免新机制嵌入复合型框架中存在的跨维度定位难题。通过逻辑上无重叠而不遗漏的二分法，结合人工智能自身的技术原理和特点，采取多层二分支式的治理框架，或将是我国人工智能法治建设的有益抉择。

［本文系国家社会科学基金一般项目“算法解释制度的体系化构建研究”（22BFX016）阶段性成果。］

① Nick McKenna, Tianyi Li, Liang Cheng, et al., “Sources of Hallucination by Large Language Models on Inference Tasks,” *In Findings of the Association for Computational Linguistics: EMNLP* 2023, pp. 2758–2774.

② 张凌寒：《中国需要一部怎样的〈人工智能法〉？——中国人工智能立法的基本逻辑与制度架构》，《法律科学（西北政法大学学报）》2024 年第 3 期。

中国人工智能法的调整对象：锚点、主客坐标与效力边界

郑志峰，西南政法大学科技法学研究院副院长、教授

人工智能科学立法的首要问题是确定其调整对象，准确把握立法调整对象背后的规律和本质。对此，笔者认为，中国人工智能法的调整对象可以从三个维度进行界定。

中国人工智能法调整对象的概念锚点

对于人工智能法来说，其调整对象的界定、制度框架的搭建以及具体规则的设计，都需要对人工智能进行科学、合理的法律定义，以便充当识别社会关系、定性法律关系的第一道“滤网”。[①]

第一，人工智能的界定。人工智能的技术逻辑是模拟、增强甚至替代人类智能，这直接导致机器智能的产生，使得社会关系的运转多了一种介入力量，直接冲击现行法律制度以人类为中心展开的预设前提。[②]而人工智能的智能特征又集中表现为自主性，即人工智能能够自主地进行计算、推理、学习、决策甚至行动，在特定的输入端与输出端之间将人类排除“环外”。据此，自动驾驶汽车、深度合成、人脸识别、生成式人工智能都符合人工智能的定义。与此相反，辅助驾驶并非自动驾驶，需要人类驾驶员处于“环内”，不符合人工智能的定义；达芬奇手术机器人完全由外科医生操控，没有人工智能的自主性；而人形机器人强调的是外形与人形相似，也不能当然地归入人工智能的范畴。

第二，人工智能的形式。明确人工智能的定义可以很大程度上划定人工智能法的调整范围，但人工智能作为一种技术，本身不能成为人工智能法的调整对象，故还需要明确人工智能法调整的客体形态。[③]对此，欧盟《人工智能法案》采取了“人工智能系统”这一概念，我国学界两部人工智能法建议稿则使用了“人工智能产品或者服务”的表述。笔者认为，“人工智能产品或者服务”一词更加适合本土立法的语境。其一，人工智能技术通常需要以产品或者服务的形式呈现在公众面前，以便进入法律视野。其二，人工智能产品与服务适用的法律规则不完全一样，如人工智能产品受产品责任约束，而人工智能服务可以适用通知删除规则等，区分两者可以实现立法的精细化。[④]其三，人工智能产品或者服务的表述能够与《产品质量法》《民法典》《刑法》《生成式人工智能服务管理暂行办法》等法律法规进行衔接，延续以往产品与服务两分的调整规则。

与此同时，当前人工智能主要是应用于特定领域、特定事项、特定任务的专用型人工智能，但不排除未来出现通用人工智能的可能。那么，人工智能法是否需要关注通用人工智能就成为一个问题。欧盟《人工智能法案》就通用人工智能模型做了专章规定，第51条规定，委员会可以根据适当的技术手段和方法，对通用人工智能的能力和影响进行评估，以便确定是否属于具有系统性风险的通用人工智能。同时，当一个通用人工智能模型用于训练的累计计算量，以浮点运算计大于10^{25}时，应推定该模型具有高影响力。国内方面，《人工智能示范法（专家建议稿）》规定了基础模型，《人工智能法（学者建议稿）》就通用人工智能

① 陈亮、张翔：《人工智能立法背景下人工智能的法律定义》，《云南社会科学》2023年第5期。
② 参见郑志峰：《人工智能的法律挑战与规制重点》，《月旦民商法杂志》2022年第6期。
③ 参见韩旭至：《人工智能法的调整范围》，《北京航空航天大学学报》（社会科学版）2024年第3期。
④ 参见郑志峰：《人工智能产品责任的立法更新》，《法律科学（西北政法大学学报）》2024年第4期。

的内涵做了明确。考虑到大模型带来的“智能涌现”现象，人工智能法应当对通用人工智能进行适当的超前部署，以增强立法的前瞻性。

第三，人工智能的例外。人工智能的概念非常宽泛，是否所有的人工智能类型都需要交由人工智能法调整，值得进一步思考。一是军事人工智能。军事人工智能直接关系国家安全，要求军事人工智能像普通人工智能产品或者服务那样履行报备批准、公开透明、审计、伦理审查等义务并不妥当。同时，军事人工智能有着特定的用途，社会公众一般无法接触，适用专门的法律法规以及相关的国际法规制更加合适。二是免费开源人工智能。免费开源意味着人工智能需要开放源代码，任何人都可以自由无偿地查看、修改和分发。免费开源人工智能对于技术创新具有促进作用，其伦理基础是社会鼓励公民自愿合作的精神。基于此，《人工智能法（学者建议稿）》规定本法不适用于“免费开源的人工智能”。然而笔者认为，免费开源人工智能并非毫无风险，一概排除在人工智能法的调整范围之外并不合适，如欧盟《人工智能法案》就将高风险的免费开源人工智能模型纳入监管。建议我国人工智能立法主动明确免费开源人工智能的界定，合理豁免相关主体的义务与责任。

中国人工智能法调整对象的主客坐标

人工智能法具有科技法的属性，其调整对象需要指向人工智能科技活动及相关主体。对此，可以从主体与客体两个方面，对人工智能法的调整对象进一步加以描述与定位。

第一，人工智能开发者与开发活动。对于人工智能法是否调整开发者与开发活动，理论和实务界存在不同的做法。欧盟《人工智能法案》采取提供者与部署者的二元主体结构，明确指出“本条例应支持创新，尊重科学自由，而不应损害研发活动。因此，有必要将专门为科学研究和开发目的而开发和提供服务的人工智能系统和模型排除在其范围之外”。国内方面，《生成式人工智能服务管理暂行办法》同样不涉及人工智能开发阶段，将不提供生成式人工智能服务的活动排除在外。与之不同的是，学界两部人工智能法建议稿都明确将开发者与开发活动纳入调整范围。

笔者认为，人工智能法应当调整开发者与开发活动。首先，开发活动是人工智能科技活动的重要组成部分，将之纳入调整对象可以实现对人工智能全生命周期的治理。其次，算法设计、数据标注、模型训练和优化等开发活动是人工智能风险产生的源头，将开发者与研发活动纳入调整范围，可以事半功倍地预防、管理人工智能的风险。最后，人工智能法应当调整的是面向市场的人工智能开发活动，纯粹的科学研究活动应当排除在外。欧盟《人工智能法案》就规定“本条例不适用于人工智能系统或模型在投放市场或提供服务前的任何研究、测试和开发活动”，但“在真实世界条件下进行的测试不在此豁免范围内”。因此，建议人工智能法采用开发而非研发一词来界定相关主体与活动。

第二，人工智能提供者与提供活动。提供活动处于开发活动与使用活动的中间阶段，对人工智能的风险治理起着承上启下的作用，是人工智能法重点调整的对象。唯有明确提供者的权利、义务与责任，方能推动人工智能的技术创新与大规模应用，真正回应人工智能产业发展的需求。对此，我国学界两部人工智能法建议稿都明确将提供者与提供活动纳入调整范围。与此同时，《新一代人工智能伦理规范》使用了“人工智能供应活动”的概念，指代人工智能产品与服务相关的生产、运营、销售等。有学者则采用了“人工智能应用活动”[①]的表述，指人工智能系统提供者向用户提供人工智能系统服务的活动。笔者认为，相较于供应活动、应用活动，提供活动一词更加合适。首先，应用活动似乎可以包括供应与使用两个环节，并不限于提供活动，

① 侯东德：《人工智能法的基本问题及制度架构》，《政法论丛》2023年第6期。

指向不够精准。其次，供应活动与提供活动两者大体相当，但供应活动更加像是一个产业链视角的术语，强调人工智能的生产、运营、销售等活动，而提供活动更具法律色彩。最后，欧盟《人工智能法案》以及我国学界两部人工智能法建议稿都使用了提供者的概念，同时《生成式人工智能服务管理暂行办法》也使用了提供者这一概念，人工智能法使用提供者的概念可以节约立法成本。

第三，人工智能使用者与使用活动。除了开发活动、提供活动外，使用活动也应当是人工智能法调整的重要对象。一方面，相较于开发者、提供者，使用者对于人工智能的风险管理有着重要的影响力。即使人工智能产品或者服务本身没有安全隐患，但若使用者没有进行合理使用甚至是滥用的话，仍然会引发严重危害。例如，有条件自动驾驶汽车需要人类用户在紧急情况予以及时接管；生成式人工智能的生成结果高度依赖用户的关键词指令；人脸识别既可以帮助企业用于安防、打卡等正常用途，也可能被滥用为监视员工的不法工具。另一方面，使用者是离人工智能产品或者服务风险最近的群体，常常是人工智能技术风险的侵害对象，理当成为人工智能法的调整对象。

立法方面，欧盟《人工智能法案》使用的是部署者（deployer）一词，而我国学界两部人工智能法建议稿则采用了使用者的概念，两者本质上是一致的。具言之，欧盟《人工智能法案》之前的版本一直采用的是使用者（user），直到2023年12月才替换为部署者（deployer），但具体定义并未改变。与此同时，使用者与用户很多时候可以等同使用，但用户更多地指向作为个体的使用者，不能涵盖企业、政府机构等组织类群体，使用者则更具包容性。此外，使用活动非常广泛，是否都属于人工智能法的调整范围不无疑问。对此，欧盟《人工智能法案》规定“在个人非职业活动中使用人工智能系统的情况除外”。国内方面，《人工智能法（学者建议稿）》也明确“自然人因个人或者家庭事务使用人工智能的”，不适用本法。笔者认为，个人或者家庭事务中的人工智能使用活动主要免除的是使用者的合规义务，并非一概不适用人工智能法。例如，开发者、提供者仍然需要保障使用者的权利，使用活动致人损害的使用者需要承担责任。

第四，人工智能监管者与监管活动。从性质上看，人工智能法具有领域法的特征，其不等于部门法，也不等于人工智能＋部门法。[①]这意味着人工智能法具有公私法交融的特征，既要调整平等主体之间因为人工智能的研发、提供与使用活动产生的法律关系，也要调整不平等主体之间围绕人工智能监管活动产生的法律关系。从内容层面来看，人工智能法既要克服规制科技不足，也要避免规制科技过度，至少应包括人工智能市场促进法、人工智能政策促进法与人工智能风险治理法的内容。其中，人工智能政策促进法、人工智能风险治理法都属于监管法律关系，是人工智能法的重要部分。

国内方面，两部人工智能法专家建议稿都明确规定调整人工智能的监管活动。例如，《人工智能示范法2.0（专家建议稿）》设有“人工智能管理制度”与“人工智能综合治理机制”专章，并规定了包括国家人工智能办公室、国家人工智能主管机关、中央人工智能领导机构、各级人工智能主管机关、其他有关部门和军队有关部门等各种监管者。《人工智能法（学者建议稿）》也设有“监督管理”专章，同时就人工智能的监管部门、监管职责以及开发者、提供者、使用者违反监管规定的行政责任承担进行了规定。欧盟《人工智能法案》更是将人工智能监管活动贯穿始终，确立了基于风险的分级规则、“全链条”的监管措施、实验性的“沙盒监管”等制度。[②]

中国人工智能法调整对象的效力边界

除了从内部界定人工智能法的调整对象之外，还需要从外部确定人工智能法调整对象的边界。这就涉

① 参见郑飞：《论人工智能法的理论体系》，《法治研究》2024年第4期。

② 参见董新义、梅贻哲：《“人工智能法总则”建构原则与理念——欧盟立法经验之镜鉴》，《数字法治》2024年第2期。

及人工智能法的空间效力及其与相关立法的关系。

第一，人工智能法的空间效力。面对国际竞争的复杂局势，人工智能法应当规定对特定的境外人工智能科技活动保有管辖权。首先，当前人工智能的国际竞争已经上升到国家安全与国家战略的高度，一些西方国家长期通过长臂管辖规则对我国人工智能产业进行打压和排挤。此种背景下，我国人工智能立法必须予以反制。其次，从以往立法实践来看，域外管辖已经成为维护国家主权的重要方式，人工智能法应延续这一做法，全方位维护我国网络主权、数据主权以及人工智能主权。最后，人工智能的风险具有跨境性，人工智能法必须要应对境外人工智能活动带来的风险，践行总体国家安全观。对此，我国学界两部人工智能法建议稿都确立了域外管辖规则，明确在境外开展人工智能活动影响或者可能影响我国国家安全、公共利益以及个人、组织的合法权益的都属于我国法律管辖范围。建议未来我国人工智能法在确定最大连接点的基础上，细化域外管辖的具体情形。

第二，人工智能法与智能要素法的协调。2020 年 8 月，美国安全和新兴技术研究中心发布的《人工智能三要素以及对国家安全战略的意义》指出：“可以用一句话来概括现代人工智能的复杂性，即机器学习系统使用计算能力来执行从数据中学习的算法。”由此可知，人工智能神奇的关键在于数据、算法及算力三要素。从这个角度来看，人工智能法需要处理好与智能要素立法的关系，协调好各自统辖规制的边界，最大程度避免重复规制与规则冲突。

一方面，人工智能法必须对智能要素有所关注，让智能要素融入人工智能的治理规则中。从产业促进角度来看，人工智能法应当鼓励算力基础设施建设、算力资源利用、算法模型创新、数据要素供给、数据合理使用等；从风险治理角度来看，数据要素、算法模型、算力资源会直接影响人工智能产品或者服务的安全、稳定、有效，人工智能法应当就训练数据质量、算法模型监管、算力资源供给等进行规定。另一方面，人工智能毕竟不同于智能要素，人工智能法应当适当留白，否则会偏离人工智能法调整的中心主线。[①]例如，人工智能的个人信息保护可以援引适用《个人信息保护法》；人工智能的数据安全可以通过《数据安全法》来解决；对于训练数据以及输出端生成数据的产权问题，也应当留给未来的数据立法。

第三，人工智能法与智能应用法的关系。对于人工智能法如何处理具体人工智能应用，国内外存在不同做法。国内方面，《人工智能示范法 2.0（专家建议稿）》并未就具体人工智能应用作出规定，仅从整体层面对人工智能进行规制。与之不同的是，《人工智能法（学者建议稿）》设置了“特殊应用场景”专章，就执法人工智能、司法人工智能、新闻人工智能、医疗人工智能、社交机器人、生物识别、自动驾驶、社会信用人工智能、通用人工智能等典型人工智能应用做了专门规定。此外，欧盟《人工智能法案》在进行人工智能分级分类时，也提到了生物识别、深度合成、司法人工智能、执法人工智能等具体的人工智能应用。

人工智能本身是一种通用赋能型技术，可以广泛应用于医疗、交通、金融、教育、司法、物流等领域。欧盟《人工智能法案》就指出：“人工智能是一个快速发展的技术族，能够为各行各业和社会活动带来广泛的经济、环境和社会效益。”实践中，不同人工智能应用的风险表征、治理机制不尽相同，人工智能法不能仅仅停留在抽象的人工智能层面，还要看到不同人工智能应用的特点，使人工智能治理规则更具针对性、操作性。与此同时，人工智能法规定具体人工智能应用也符合总分立法的逻辑，可以实现人工智能一般规则与特别规则的融合，为未来智能应用的专门立法提供指导。基于此，人工智能法应当对具体人工智能应用的典型挑战进行适当规制，如自动驾驶汽车的电车难题、医疗人工智能的辅助定位、司法人工智能的公平透明等。

① 参见郑志峰：《人工智能法的表白、留白和补白》，《北京航空航天大学学报》（社会科学版）2024 年第 3 期。

人工智能监管中的数据治理层次论

许身健，中国政法大学数字社会治理研究院院长、教授

随着人工智能技术的迅猛发展，数据作为其核心驱动力的重要性愈发凸显。无论是机器学习模型的训练，还是智能系统的决策制定，都离不开大量、高质量数据的支持。然而，数据的收集、存储、处理和使用过程涉及众多法律、伦理及社会问题，这些问题的解决直接关系到人工智能技术的健康发展和应用前景。[①] 数据治理作为人工智能应用的基础和保障，其重要性不言而喻。

数据治理涵盖了数据从诞生到消亡的整个生命周期，包括数据的采集、存储、管理、分析、共享和销毁等各个环节。它不仅要求对数据进行有序、高效的管理，还强调数据的质量和安全性，以及数据的合规使用。[②]在人工智能领域，数据治理更是关乎算法的准确性、模型的可靠性以及最终决策的公正性。没有良好的数据治理，人工智能技术就可能陷入数据混乱、质量低下、安全隐患重重的境地，甚至可能引发严重的法律、伦理和社会风险。

在人工智能监管过程中，数据治理应当被赋予重要地位，并作为底层逻辑体现在人工智能产品和服务中。对此，利益相关者应当在价值层面形成共识，共同建构完善的数据伦理体系；人工智能供应链中的参与者应当建立和完善内部的数据治理体系，推动数据治理技术和方法的创新与发展；立法者和政策制定者应当通过制定相关法律法规及政策等，明确数据的权属、使用范围、安全保护措施等关键问题，为人工智能领域的数据治理提供法律依据和指导原则。

价值层次：建立完善的数据伦理体系

对于数据采集、分析、加工等一系列信息技术而言，承载着个人情感、态度、价值观的个人信息乃至反映社区、城市居民工作动向、生活水平的民生发展的数据，都不过是计算机系统中一串“0”“1”交织的编码，并不具有任何道德意味，数据技术具有价值中立的工具属性。然而，为了使数据功能得以运用，使数据价值得到展现，在数据的收集、处理和使用过程中，却不可避免地涉及一系列道德与价值判断。人工智能时代的新问题不免让我们思虑：未来需要什么样的数据环境？应该期待什么样的数据流转？利用数据的标准和底线何在？这些问题恰恰都统一于良善的数据伦理体系之中，完善人工智能监管中的数据治理，首先应在价值层面建立与法律法规相匹配的数据伦理框架及体系，[③]这是数据治理的基石。

对数据的获取与持有是对数据进行资源整合和加工的先决条件，也是进行人工智能技术创新的基础。为此，数字社会已经形成了在用户知情同意基础上的数据收集机制，但随着数据应用领域扩大、程度加深、场景多元，对用户信息的收集边界与幅度也不断扩张，个人的姓名、年龄、住址、职业、家庭等基础信息和消费习惯、性格喜好、生活方式等个性化信息都不断被纳入收集的范围。超级数字平台的崛起更是改变了知情同意的原始样态，其通过“不同意即拒绝服务”，[④]让用户“心甘情愿”又源源不断地供给个人信息。“必要数据”语焉不详的法律界定和实务判断令数据收集

① 参见张涛：《生成式人工智能训练数据集的法律风险与包容审慎规制》，《比较法研究》2024年第4期。

② 参见陈刚主编：《数据资源规划与管理实践》，北京：北京交通大学出版社，2021年，第150—151页。

③ 参见大卫·马滕斯：《数据科学伦理：概念、技术和警世故事》，张玉亮、单娜娜译，北京：中国科学技术出版社，2024年，第5—7页。

④ 刘权：《网络平台的公共性及其实现——以电商平台的法律规制为视角》，《法学研究》2020年第2期。

的窗口增加、挖掘数据信息的机会增加。同时，这也意味着处于数据汪洋中的普通个人的隐私也存在被暴露的风险。《全球人工智能治理倡议》提出“以人为本”的理念，强调以增进人类共同福祉为目标，以保障社会安全、尊重人类权益为前提，确保人工智能始终朝着有利于人类文明进步的方向发展。具体到数据治理领域，数据伦理是以人为本位的伦理，人的权利与价值处于中心位置，这便要求在开展数据采集和利用活动时切实保护个人隐私。

数据治理环境中的信息不对称、权力不平等等问题导致人的自主性受到限制，人工智能开发中的数据处理活动应当坚守公平的知情同意原则。有别于传统产业发展中生产主体与消费主体间的单向联系，数据视阈下的供给与需求关系发生了双向乃至多向的变化。基于数据的易复制、易传播、价值不易损耗等特性，数据要素可以反复、自由、便捷地在不同市场主体之间发生交换与流转，原先的消费者将不仅仅是被动地提出需求、接受供给的主体。通过参与市场分配产生的个人信息、消费数据本身也成为数据资源的一部分，消费者同时也成为支持新的生产—消费活动的数据的持有人，从而成为持有原始数据，可以进行数据授权加工，分享数据收益的小型数据生产主体。这种单向关系的突破，意味着数据持有者、平台方要改变既往权利义务分配不对等的“要挟式”格式条款，不以简单的拒绝使用而造成用户个人客观上“被知情同意”的窘境。

同时，要规范利用数据匿名化等技术手段处理涉及原始信息主体的敏感数据，对非经脱敏处理的敏感数据限制进行流转与运用。除坚持数据采集时对个人隐私的保护之外，还应探索数据资源使用后对相关敏感数据的处置机制，个人隐私“被删除”与“被遗忘”的权利同样值得关注，[①]对原始信息主体的个人隐私的保护应该覆盖数据的全流转环节与全生命周期。

数据的聚合为人工智能的技术飞跃提供了跳板，同时也可能引发人的主体性困境，数据治理应当始终坚持以人为本的价值观。随着人工智能技术的进一步发展与广泛应用，数据分析的能力得到进一步提升，主体基于持有的足够丰富的信息，将能够作出足够精确的行为预判。通过行为预判以及此种预判与分析的叠加，将不断定向构筑更加精确的“信息茧房”。社会对数据服务的期待越来越多，依赖数据的需求从简单而变得逐渐复杂，人的生存环境正发生数据化的演进。随着数据量级的膨胀与倍增，人对于数据或技术的信任与依赖也与日俱增，在复杂的情境下，甚至可能“迫使”人们必须完全依赖数据的预测和结论才能作出最终决定。数据与人的主体关系发生了异位，权力化的数据带来的对人的自由意志的垄断应当引起必要警惕。对此，要确定数据使用的伦理界限，对应该训练、培育、鼓励和体现原创性思想和独特的创造性领域，要避免对数据的过度依赖和“唯数据论”的判断。数据治理最终应归于人对数据的治理而非数据对人的调控，数据不应替代思考，数据也不应垄断创造。当出现对数据及技术的利用空间无度扩大并假以先进之名时，应当坚守人的主体性，人应该在人工智能时代保留自主思虑的园地。

此外，人工智能时代的“大数据杀熟”“算法歧视”等问题屡见不鲜，这要求人工智能监管中的数据治理亟待回归促进社会公平正义的价值目标。在“算法霸权”时代，[②]向数据平台供应其赖以生存的数据要素的个体，反而受到这些庞大数据富集主体的调控与支配。从数据资本主义到平台资本主义，从个体信息资源的“剥削”到利用数据机会的“剥夺”，[③]不公平的数据使用和收益机制正令传统的“数字鸿沟”以新的形态不断涌现。正如《关于构建数据基础制度更好发挥数据要素作用的意见》（以下简称《数据二十条》）所提倡的那样，数据的使用应当体现“共同使用、共享收益”的精神，要促进数据使用的公平与透明。将增强数据流动活力与提升数据分享意愿统一起来，在注重体现按市场决定资源配

① 参见刘文杰：《被遗忘权：传统元素、新语境与利益衡量》，《法学研究》2018年第2期。

② 凯西·奥尼尔：《算法霸权》，马青玲译，北京：中信出版集团，2018年，第73—74页。

③ FAGIOLI A., “To Exploit and Dispossess: The Twofold Logic of Platform Capitalism,” *Work Organisation, Labour & Globalisation*, vol.15, no.1, 2021.

置、按劳动贡献决定价值分配的规则设计外，还应特别关注数据所承托的公共利益与社会公众对其加以访问、使用的合理期待，真正做到使数字经济的发展红利为全体人民共享，进而实现促进社会公平正义的目标。

技术层次：推动数据技术的负责任创新

在人工智能时代，数据治理技术层次的不断提升已成为推动数据技术负责任创新的关键因素。数据治理作为一种综合性的治理框架，涉及数据的质量控制、安全性、隐私保护以及合规性等方面，其核心宗旨在于确保数据的有效管理与利用，同时维护数据主体的权益和公众利益。从理论层面来看，数据治理技术的层次化发展，体现了对数据生命周期全过程的深入理解和精准把控。这包括数据采集、存储、处理、分析、共享和销毁等各环节的技术治理措施，旨在通过标准化和规范化的方法，提升数据的透明度和可追溯性，从而降低数据使用过程中的风险。负责任的创新是数据治理技术发展的重要导向。这种创新不仅是技术上的进步，更体现着对社会价值、伦理原则和法律法规的深刻考量。在人工智能发展过程中，应当积极推动数据技术创新，坚持以人为本的原则，充分考虑数据治理对个人隐私、数据安全和社会公正的影响，确保技术进步不以牺牲基本人权和公共利益为代价。

首先，数据的准确性和完整性是数据技术负责任创新的基础。在技术层面，人工智能供应链中的参与者应当建立健全的数据质量管理体系，包括数据清洗、验证和维护等举措。该体系需覆盖数据生命周期的各个阶段，包括数据的生成、存储、使用、维护以及销毁等。其中，数据清洗、验证和维护是数据质量管理体系的关键组成部分。数据清洗，涉及对数据进行审查、纠正错误和消除冗余信息的过程，目的是提升数据的一致性和可靠性。数据验证则是确认数据是否符合既定的质量标准，并保障数据在各个处理阶段的准确性。数据维护则关乎持续监控数据的健康状况，确保数据的时效性和相关性，防止数据退化和过时。通过建立严格的数据质量管理体系，我们能够为数据治理提供坚实的基础，推动数据价值的最大化，从而促进人工智能技术的可持续健康发展。

其次，数据的安全性是数据技术负责任创新的重要内容。《数据安全法》《网络安全法》《网络数据安全管理条例》等相关法律法规对数据安全保障与安全机制建设作出了明确规定。人工智能供应链中的参与者应当树立必要的数据安全意识，以机密性（confidentiality）、完整性（integrity）、可用性（availability）原则做好数据安全保障，保证敏感数据只能被授权人员访问，且数据不被篡改，能够被及时可靠地访问并留有安全冗余。在数据访问控制技术层面，应加强认证与授权管理，配置得当的访问控制列表与策略，做好用户管理，并建立定期的记录审查。在数据保护技术上，注重加密、脱敏、掩码、备份等技术储备，提升抵御网络攻击的能力。在恶意软件防护上，通过定期更新和补丁管理，定期扫描和清理，做好防病毒与反恶意软件的检查。同时，应当及时跟进适应人工智能时代的新技术应用的发展态势，建立数据安全信息特征库，优化算法等技术模型，建立数据安全风险的预警预防和识别监测机制。

最后，隐私保护是数据技术负责任创新的重要方向。在人工智能开发过程中，隐私保护已然成为最受瞩目的议题之一，其重要性在政产学研各界均得到普遍认可。在技术层面，人工智能供应链中的参与者应当积极开发隐私增强保护技术。例如，匿名化技术通过移除或替换个人识别信息，如姓名、地址、身份证号等，使得数据在保留其分析价值的同时，个人身份无法被直接识别。差分隐私则是一种更为先进的隐私增强保护技术，它通过在数据发布中添加随机噪声，

保证即使在发布聚合信息的情况下，单个数据项的泄露风险也被最小化。这些技术的应用，使得研发主体可以在不暴露个人信息的前提下对数据进行有效的分析和研究。随着人工智能技术的不断进步，新兴的隐私计算技术亦为数据隐私保护提供了新的途径。隐私计算通常包括同态加密、安全多方计算（SMPC）、零知识证明等方法，它们允许在不暴露原始数据内容的条件下进行数据处理和分析。例如，同态加密技术允许对加密后的数据进行运算，而结果仍然保持加密状态，从而可在不解密的情况下利用数据进行计算。

规范层次：迈向数据法律的整体性治理

数据具有的非损耗性、流动性、开放性、易复制传播性、非排他性、非竞争性等特点决定了数据作为生产要素之“新”，[①]也相应决定了与之相配套的“新”治理规则和治理框架的要求。传统的治理规则与框架已经不能很好适应数据引领下的发展实践，面对不够理想的驱动技术的创新实践，不够公正的数字化发展的资源利益分配，以及不够规范的数据资源的权力管控等问题，面向人工智能时代的数据治理的整体性框架和规范亟待建立。人工智能监管中的数据治理要落到实处，必须形成规则治理，做到有章可循。数据的流转运行环环相扣，对其开展综合治理亦是节点繁多，面对这一庞大的系统工程，需要有全面的数据治理法律支撑。在人工智能立法讨论如火如荼的时代背景下，作为人工智能综合治理的一个切面，要将数据治理纳入法治轨道，坚持法治引领治理，推进高质量的数据法治建设。

作为我国数据发展与治理领域的基础性政策文件，《数据二十条》具有重要指导意义。《数据二十条》指出应对数字资源持有权、数据加工使用权、数据产品经营权予以结构性分置。数据资源领域的“三权分置”与数据治理的法治逻辑具有一致性，数据法治的目的就是要建立明晰的数据产权制度、合理的数据流通制度以及公正的数据权益分配制度。遍观既往立法实践，对这三个领域并没有专门且集中的法律条文与规定，而是散见于不同的部门法中。如在数据资源确权领域，仅有《著作权法》《专利法》等知识产权领域的法律规范能够为确权提供大致的、较为模糊的参照；在数据加工、流通领域，法律规定多集中于电子商务这一具象载体，如《反不正当竞争法》《电子商务法》的相关规定；在数据产品经营及收益分配领域，《反垄断法》《消费者权益保护法》等起到主要规范作用。另外，《网络安全法》《数据安全法》《个人信息保护法》等法律规范为数据安全、数据信息处理活动领域奠定了重要基础。但是，这批已有的涉及数据领域的立法准备更多关注的是数据在生产活动中的零星或个别环节，只是将数据活动置于相对静态的环境，而并未贯通数据动态流通的全阶段、全环节，使数据运行留有法律空白。加快完善覆盖数据经济活动全过程的综合性的法律规范体系，注重数据治理领域立法的全面与综合是数据法治的发展趋势。在比较法视野下，欧盟制定了《数据法》和《数据治理法》，二者均适用于个人数据和非个人数据，并以促进数据的共享和再利用为理念，以建立欧盟数据单一市场为目标，这可以为我国的综合性数据立法提供参考与借鉴。

在数据生产阶段，数据关系及数据利益关系交织、聚合，最为关键的便是在《数据二十条》所提出的“分类分级确权”的基础上制定合理确权的法律规范，以平衡利益、提高效率、化解冲突作为确权规范的基本原则；在数据流通环节，要注重数据交易秩序的维护，确保数据交易安全、便捷、可信，制定数据交易的相关法律规范；在数据收益环节，除了坚持体现按贡献度分配的市场原则外，还要特别注重社会公共利益的保护与原始数据主体的利益衡平，科学配置“共享收益”的机制与方式。并且，要通过相关制度规范，切实保证与促进数据的可利用、可获得，让全体人民共享数

① 时建中：《数据概念的解构与数据法律制度的构建兼论数据法学的学科内涵与体系》，《中外法学》2023年第1期。

字经济的发展红利。此外，还应该配套建立适应数字社会、数据治理的数字化的权利救济与争议解决机制，通过法律规范的指引作用，带动司法机关主动以数据赋能的方式，推进数字检务、智慧法院、司法大数据等建设，以数据新动能助力司法新能动，增强司法机关在人工智能时代产出优质司法公共产品的能力与水平。

同全面的数据治理法律配套的是有效的数据治理机制。在产业数字化与数字产业化的双向发展之中，多个市场主体的联动配合及行业上下游之间的共同发力成为常态。在数据流通及数据产业链的发展中，数据市场的交易者、参与者以及相应的监管部门之间具有密切的互动关系，并且因为数据流通的种种特点，使得置身于数据交易市场的各主体的身份还发生着交错与重叠。这便对建立跨行业、跨组织、跨领域、跨部门的协同治理机制提出了新要求。笔者在实践中观察发现，我国数据治理已经初步形成了政府主管部门、企事业单位、社会组织、个人生产与开发者通力合作，多元参与的治理格局。国家数据局自 2023 年末成立以来，北京、上海、深圳等地均已积极探索跨部门协同的监管策略的落地与实施，协同配合的数据治理机制初步形成。

当然，在治理机制的探索上，各地仍存在诸多步调不一致、思想不统一的地方，如近年来各地数据交易中心的“野蛮”生长与实际投放使用的“零星”形成鲜明对比。在规则定向不清晰，政策指导不明确的时候进行的数据交易机构的重复、个别建设都在一定程度上为监管与规制的实效带来了不利影响。在提高数据要素市场化配置能力与效率，促推数据要素全国统一大市场建设的征途上，共享有活力、交易有秩序、监管有保障的有效有为的数据治理机制建设任重道远。

结语

人工智能监管中的数据治理层次丰富多样，每个层次都有其独特的角色和实现路径。数据治理不仅是一个技术问题，它涉及社会、法律和伦理等多个维度，只有通过全面、系统的数据治理体系建设和实施，才能确保人工智能技术的合法合规运行和持续健康发展。在当前和未来的人工智能发展中，数据治理将面临更多的挑战和机遇。因此，我们需要不断加强对数据治理的理论研究和实践探索，为人工智能技术的未来发展提供更加坚实的基础和保障。

在理论研究方面，数据治理的模型和工具是研究的重点。一是要构建科学的数据治理模型。数据治理模型可以帮助我们系统性地理解数据治理的各个环节和要素，包括数据收集、数据存储、数据处理、数据共享和数据安全等。通过构建数据治理模型，我们可以更好地理解各个环节之间的关系，并找到优化和改进的方案。二是要进一步开发完善数据治理的工具。数据治理工具可以帮助我们提高数据治理的效率。目前，已有一些成熟的数据治理工具，如数据质量管理工具、数据安全管理工具和数据生命周期管理工具等。但随着人工智能技术的发展，数据治理工具也需要不断更新，以应对新的需求和挑战。

在实践探索方面，我们需要通过具体的应用案例和项目，验证和优化数据治理的方法和技术。在实际的应用过程中，可能发现理论研究中未曾预料到的问题，要通过实践探索找到解决这些问题的方法。一方面，可以开展跨行业的数据治理合作项目。不同的行业在数据治理方面有不同的需求和挑战，通过跨行业合作，我们可以共享数据治理的经验和资源，提高数据治理的整体水平。另一方面，可以推动数据治理的标准化建设。标准化是提高数据治理效率和效果的重要手段。通过制定和推广数据治理的标准，我们可以规范数据治理的过程和方法，提高数据治理的透明度和可操作性。

未来，我们还需要进一步加强数据治理的国际交流与合作，学习和借鉴国际上优秀的经验和做法，不断提升我国数据治理的整体水平，为我国人工智能技术的健康发展注入新的动力。

中国人工智能立法的价值基础与伦理治理模式

张龑，中国人民大学法学院副院长、教授

人工智能不仅是一种替代劳动的自动化技术与生产工具，同时构成了新质生产力基本内涵中的劳动力要素。作为劳动力的一部分，人工智能本身有着一般生产工具所不具备的伦理属性，在创新生产方式的同时，也必然深刻改变人类的价值观与生活方式。当下，深入探讨人工智能立法的价值基础，结合中华优秀传统文化，建立符合人工智能特质的伦理治理模式，是我国人工智能法治化建设的重要任务。

人工智能立法的时代背景与价值基础

1956 年“人工智能”概念首次在美国达特茅斯学院被提出。自此之后，人工智能成为未来科技的代名词。但时至今日，人工智能都难以构成具有严格内涵与外延的术语，其泛指涵盖一切运用机器模拟人脑智慧进行思考和决策的理论、方法和技术。2023 年初，OpenAI 公司开发的 ChatGPT-4 横空出世。此后，Sora、Claude、Gemini、文心一言等生成式人工智能大模型层出不穷，引发新一轮人工智能发展高潮，朝着通用化的终极目标发起冲刺。不同于先前取得成功的各种人工智能系统，通用人工智能的目标要求机器应当具备“智能体在各种环境中实现目标的能力”。在这一席卷全球的浪潮下，人工智能技术不断在传统场景落地与应用，与传统领域融合发展，对社会生活产生越来越广泛的影响。相应地，社会对人工智能的规范与治理的需求也愈发迫切。

当前人工智能快速发展，涉及内容生成、自动决策、面部识别以及大数据分析等技术的应用不断增多，引发了社会关于数据隐私、公平性、透明度、安全和责任等问题的担忧。这不仅反映了数字社会中传统法律控制对利用人工智能等技术手段实施违法犯罪失效的困境，更使人类认识到自身面临的是一场对“人本主义”价值观的革命。[①]一系列道德和伦理挑战接踵而至，如人的自主性弱化、算法歧视与虚假信息导致的信任受损等。可见，人工智能带来的风险是托马斯·库恩所言的范式革命的风险，立法构建监管体系和新型伦理治理机制，是回应人工智能的范式

挑战，确保社会有序转型的必经之路。

人工智能发展从一开始就是全球性的，引发国际竞争势所必然。各国的监管机构在本国治理经验的基础上，纷纷致力于将自己的方案发展为全球共识，从而引发了一场关于人工智能治理的全球竞争。哪些国家能够率先制定并实施有效的治理框架，将直接影响其在人工智能治理领域的政策输出能力和国际影响力，也会为本国人工智能技术出海提供便利和支持。特别是，个别西方霸权国家将自身价值观上升为技术标准，不可避免地威胁到世界文化的多样性。目前来看，我国在人工智能发展的很多方面居于世界领先地位，这决定了我国人工智能治理不仅要开展广泛的国际合作，还要冷静应对激烈的竞争。因此，我国应当积极发挥引领全球人工智能治理的作用，结合本国立法实践，提出中国方案，贡献中国智慧。

当前我国人工智能领域呈现出快速发展的趋势，百度、阿里、科大讯飞、华为等科技巨头纷纷下场，使得对人工智能的

① 参见齐劲沣：《人工智能的法理定位与风险规制》，《数字法治评论》2023 年第 2 期。

立法监管需求变得愈发紧迫。然而，人工智能法律监管的力度与市场活力之间具有复杂的相关性。一般来说，监管力度弱则市场活力强，这可以从我国互联网产业的发展史中得到印证。过于严格的监管措施可能会让市场主体在选择研究方向时变得更加保守，导致一些可能具有引领性的技术因为监管的不确定性而无法得到充分探索和发展，引发“不发展就是最大的安全”的悖论。但是，人工智能的一系列特质又决定了放松监管会带来较大的国家和社会安全隐患。特别是，全球范围内的大模型竞赛使得人工智能立法对监管力度的把握变得非常微妙。有观点认为，哪个国家掌握了最先进的人工智能，谁就拥有了选择“对齐”[①]哪种人类价值观的权力。最新出台的欧盟《人工智能法案》试图以风险分级的模式降低人工智能企业的义务，即便如此，仍有超过 150 位欧洲企业高管持反对意见并签署公开信，认为该立法案将危及欧洲在大模型领域的竞争力和技术主权，无法有效应对来自国际层面的挑战。因此，在“不发展就是最大的不安全”成为社会共识的前提下，人工智能立法监管在价值基础上必须从人民的根本福祉出发，实现安全与创新之间的平衡。

我国立法始终以人民为中心。人民是现代性观念，作为启蒙时代的产物，它的一个重要意涵在于，是人而非上帝成为立法者。因此，以人民为中心包含了以人为本的精神。维护人工智能的技术安全与促进技术创新是以人民为中心这一理念下的两个子原则，二者辩证统一、相辅相成。维护人工智能技术安全保障的不仅是短期的技术应用利益，更是鼓励创新和可持续发展的长期赋能效益。为了找到技术安全和技术创新的平衡点，我国的人工智能立法应当成为一套具有时间曲线的法律体系。在人工智能技术开拓创新时期，立法要充分考虑技术发展的不确定性，在不触及国家和社会根本安全底线的基础上，政府应承担促进义务，使企业获得更多的自主权。在技术进入迭代更新时期，立法应注重安定性、融贯性，筑牢各领域安全的篱笆，向保障消费者合法权益倾斜。在技术进入成熟期和缓慢成长阶段，立法应注意维护市场公平竞争，限制和反对垄断，为培育和孵化新技术创造生态。

告别意志论的养成主义伦理治理模式

为应对安全与创新平衡的挑战，立法监管面临着一种新型复杂局面，伦理、法律和技术等多种规制理念与手段需要交叉综合运用。无论是立法规制，还是设定相应的技术规范，伦理治理的思路都必须被融合其中。然而，数字人本主义固然要遵循，但如前述，在西方传统里，人民替代上帝成为立法者，人民概念里包含着很强的意志论思维，在人–机共生的中国语境下，需要有新的伦理治理思路。我国《新一代人工智能伦理规范》规定：“将伦理道德融入人工智能全生命周期。”这一表述意味着，我国采用了新型伦理治理模式，即一种贯穿于人工智能的训练、形成到运用全过程的“数据—向善”“缺失—填平”以及“失范—纠正”[②]的养成主义伦理模式。

在新一轮的图灵测试中，ChatGPT-4 的表现虽然与人类仍有辨识上的差距，但其通过率已高达 54%，[③]这种模仿游戏已经在一定程度上证明其拟人化工作的成功。美国哲学家约翰·塞尔在 20 世纪 80 年代通过“中文房间”思想实验（the Chinese Room Argument）就对此提出过质疑，[④]他认为，人工智能在本质上与人是有差距的，人工智能不等于也不能完全替代人。但是，人工智能科技的飞速发展与运用，充分表明以人工智能为代表的硅基生命在存在论意义上具有自身独特的代具文化。若希望维持人–机的共同存续与良性互动，人类固有的伦理价值必须被保护，硅基生命的代具文化按照人的伦理来塑造，“数字人本主义”[⑤]的伦理理念必须得到遵循。

从技术路径上看，生成式人工智能主要走的是以

① 王焕超：《直面“价值对齐”挑战》，《瞭望》2024 年第 26 期。

② 张龑：《例外状态与文化法治国》，《法学家》2021 年第 4 期。

③ Cameron R. Jones, Benjamin K. Bergen, “People cannot distinguish GPT-4 from a human in a Turing test,” https://arxiv.org/pdf/2405.08007v1, 2024 年 6 月 24 日访问。

④ 参见塞尔：《心灵、大脑与程序》，玛格丽特·A. 博登主编：《人工智能哲学》，刘西瑞、王汉琦译，上海：上海译文出版社，2006 年，第 73—78 页。

⑤ 程乐：《“数字人本主义”视域下的通用人工智能规制鉴衡》，《政法论丛》2014 年第 3 期。

贝叶斯优化的经验概率为主的路径。面对生成式人工智能，一些人无意识地采用了意志论下的生成主义治理模式，即认为当人工智能软件或硬件制造出来之后，再对其进行人文主义约束，此种控制理论表面上希望为人类与机器一起生活的伦理方式提供一个场域，实际上是增设了一个机器赛道与人并驾齐驱，导致了人机对抗。事实上，这种意志论生成主义伦理模式与生成式人工智能的经验概率逻辑并不相符，与后者相适应的是人机互动的新生产模式，它需要的是一种可突破单纯形式法治和决策型技术主义路线的新伦理治理方案。[①]据此，告别意志决断论的生成主义思维，结合中华优秀传统文化的养成主义伦理模式更符合贝叶斯经验概率的路径。[②]也就是说，对于生成式人工智能来说，数据集、训练机制以及影响训练者本身的人文环境及其伦理内涵成为贝叶斯优化中高概率的语言文化组合，对人工智能的伦理养成产生关键性作用，可以形象地称之为人工智能机器的“伦理养成之家”。[③]

当前来看，立法监管所力求避免的伦理“缺失”与“失范”，实际上是人工智能从数据训练到技术应用无意识地遵循意志论的生成主义思维所致。比如，人工智能系统常被设计为具有自主决策的能力，而缺乏伦理规范就可能使其做出错误的判断或决策，对人类利益造成严重的不可逆损害；使用人工智能进行数据收集会增加数据泄露和滥用的风险，侵犯个人隐私权；算法标记的路径依赖和“大数据杀熟”导致人工智能系统存有性别、地域、种族等偏见，侵犯人类的平等价值与尊严。一些学者就此提出“填平”“纠正”的解决方案，目的是为了确保人工智能的工作目标和行动与人类价值观保持一致，也就是实现“价值对齐”。[④]其中，人类价值观的实际内容还在于设计者所输出的意图和目标，即“设计者伦理品质”问题，这就涉及跨国家、跨地域和跨文化等问题，这都要求从意志论生成主义的思维束缚中解放出来，开辟顺应中华文化和伦理的养成主义伦理治理模式。确切地说，即使人工智能能够以简单博弈等方式充分再现自然选择过程，满足形成的价值系统生成的进化条件，它充其量也只是具有某种特殊偏好的“道德主体”。[⑤]无论是发展人工智能自身的代具文化，还是按照人的主体性塑造人工智能，都意味着摒弃意志论生成主义的科技产品思维，法理上则是跳出法律与道德二分法的思维定式，对人工智能的生产者、人工智能的数据内容以及人工智能的行动加以规制的法律，应直接体现人类自身的伦理价值，以“人文化成”的养成模式教化出具有人伦特质的人工智能。

立法中养成主义伦理治理的制度内涵

从世界范围看，尽管人们已就技术伦理与社会伦理的技术实现达成了诸多共识，但基于人工智能的特质将既有的科技伦理框架转化为有效的法律治理方案尚存争议。在英国学者罗杰·布朗斯沃德看来，进入法律 3.0 时代，技术主义框架下的规制手段将与融贯主义的法教义学、规制工具主义的法政策学共存，形成一种混合法律思维状态。为完成以预测和事前防止替代事后的惩罚、补偿或恢复，法律监管必然愈发具有技术主义的色彩。表面上看，这种设想是强调提前介入强化监管，实则表达了人工智能的特质，即它不仅是一种劳动工具要素，还是一种劳动者要素。如前所述，当人工智能作为劳动者或劳动者要素参与到人类生产和生活当中，国家立法采取养成主义的伦理治理模式就至关重要。

早在 2019 年，国家新一代人工智能治理专业委员会就发布了《新一代人工智能治理原则——发展负责任的人工智能》，提出八项原则。随后，《网络安全标准实践指南——人工智能伦理安全风险防范指引》《新一代人工智能伦理规范》（2021）和《生成式人工智能服务管理暂行办法》也相继出台。从内容上看，这一系列文件已经初步展示出我国就人工智能治理所采取的基

① 参见张龑：《例外状态与文化法治国》，《法学家》2021 年第 4 期。

② 参见张龑：《用“人文化成”模式治理 ChatGPT》，《法治周末》2023 年 5 月 11 日。

③ 张龑：《论我国法律体系中的家与个体自由原则》，《中外法学》2013 年第 4 期。

④ Ariel Conn, “How Do We Align Artificial Intelligence with Human Values?,” https://futureoflife.org/ai/align-artificial-intelligence-with-human-values/, 2024 年 6 月 24 日访问。

⑤ 温德尔·瓦拉赫、科林·艾伦：《道德机器：如何让机器人明辨是非》，王小红等译，北京：北京大学出版社，2017 年，第 85—92 页。

本模式以及提出的伦理原则方面的要求。当然，我国目前还未制定出一部完整系统的《人工智能基本法》，所以还没有综合性的国家制定法对人工智能伦理治理模式给予权威的选择与确认。可是，这里也包含一个悖论，即在人为自己立法的范式框架内，伦理被认为与法律不同，伦理只是作为判断法律之道德性和合理性的标准而作为论证理由使用。[①]故强调科技伦理先行是积累人工智能实践的社会经验，为立法选择与设定伦理监管模式提供先期尝试。实则不限于此，科技伦理先行之外，需要重新定位人定法与人–机共生的伦理之间的结构关系，从而达成新的伦理共识，制定出符合人工智能本质属性的人工智能基础立法。为此，中国社科院法学所牵头起草的《人工智能示范法 2.0（专家建议稿）》做了一次极为有益的尝试，它结合中华优秀传统文化，采用了中国的伦理养成主义治理模式，构造了一个新型法律与伦理的关系框架，具体体现为：

建议稿将诸如人文主义、人机共生、技术安全等伦理治理原则贯穿于立法的结构、总则与具体内容之中，使得人工智能在每个主要环节都得到伦理规范的塑造，养成为一个具有人文伦理素养的劳动主体或主体要素。首先，在结构上，《人工智能示范法 2.0（专家建议稿）》采取了总则—分则模式，总则部分主要是明确了立法目的与治理原则，使得分则的具体规定在实质的伦理内容上有所附丽；分则结合不同责任义务主体和发展人工智能的流程为伦理治理的各个环节提供规则支撑，并为应对未来的发展变化预留出足够的空间。其次，在总则里，立法依据、适用范围与总的治理原则点明了该法以人民为中心、坚持安全与创新的主旨，以人为本原则表明该法秉持将人工智能的整个生命周期处于人类控制与监督之下的理念，而后这些主旨和理念所凝聚的伦理共识贯穿于第 5 条到第 13 条的原则之中，特别是第 14 条明确规定：“从事人工智能研发、提供和使用活动应当……尊重社会公德和伦理道德”。最后，在分则里，从国家对发展人工智能的支持与促进，到人工智能的研发，再到人工智能的综合治理，都遵循着伦理养成主义的基本思路。如在支持和促进部分，包括数据要素供给（第 18 条第 2 款），专业人才培养（第 20 条）、财政支持（第 21 条）、人工智能特区建设和授权立法（第 24 条）以及负面清单设置（第 25 条—第 33 条）等，都是意在构建整体良好的人文环境和生态；在研发部分，无论是与安全性相关的义务（第 34、35 条），还是审计义务（第 36 条）和风险管理义务（第 41 条），都与数据、语料和模型等要素的审查监管有关，伦理合规性是其中一个重要判断标准。建议稿采取了有效措施确保可以随时采取介入、接管等措施，以避免歧视和偏见。对于人工智能研发者，更是强调了数据投喂和模型建设等应遵守的伦理原则。在综合治理方面，建议稿特别强调了伦理审查义务（第 42 条），并设置了专门的伦理审查委员会（第 55 条）。可见，从总则开始到第五章的整个治理方案，正类似于国家和监护人对未成年人的养成和监管，形成了一个“环境营造—风险防控—习惯养成—责任追究”的养成主义伦理治理链条。

当然，为了更加完整地贯彻伦理养成主义的治理模式，塑造出一个更加完整的法律与伦理的新制度结构，还可在建议稿中增设一些章节与条款。第一，可在人工智能立法中专章设置“充分尊重和保护人权”，要求项下的各项基本权利（主要针对使用者），也即扩大《人工智能示范法 2.0（专家建议稿）》第 14 条的保护范围，确保技术的发展不损害人类的尊严和权益，结成一条不可触碰的保护绝对价值的红线。第二，在相应的章节可适当增设人工智能使用者的义务，进一步明确研发者—使用者—提供者这一三层主体共生的伦理结构，加强对使用者身份的核实，实施实名制，避免其采用不符合伦理要求的使用方式。第三，可用劳动权益保障的思路处理人工智能的数据处理和标注等问题，确保人工智能沿着造福于人类公共福祉的方向有序、高效、可持续发展。

① 参见田静：《数字经济时代人工智能立法的后设伦理学规范检省》，《社会科学论坛》2023 年第 1 期。

人工智能伦理审查制度的立法思路

李学尧，上海交通大学凯原法学院教授、
法律与认知智能实验室主任

人工智能伦理与生物医学伦理具有对称性、延续性的关系，但人工智能伦理的责任主体、审查对象，及实质性原则和规则（如自主原则）的问题意识及其权重都有异于生物医学伦理。生物医学伦理审查过程体现为“大众道德直觉”与“无法禁止的技术创新”之间的妥协，人工智能伦理审查仍有此性质，但其“大众道德直觉”可以通过技术途经直接落实。人工智能的伦理挑战主要集中于部署和应用阶段，而生物医学伦理主要聚焦于研发和开发阶段。由上述分析而得，人工智能伦理审查制度宜在科技伦理治理框架内单行立法，并在伦理审查责任主体、审查启动条件、专家构成、议事规则以及审查结论的法律效力等方面采取有异于生物医学伦理的立法思路。

从生物医学伦理制度寻求启示

2022 年国家网信办颁布的《互联网信息服务算法推荐管理规定》第 7 条，是我国立法意义上构筑人工智能伦理审查制度的开端。2023 年 7 月，国家网信办等 7 部委颁布了《生成式人工智能服务管理暂行办法》（以下简称《暂行办法》）则在实体法规则意义上例举了人工智能伦理的相关原则。随后，科技部等单位还颁布了具有规范性文件性质的《科技伦理审查办法（试行）》（以下简称《伦理审查办法》）并在附录“需要开展伦理审查复核的科技活动清单”中例举了必须开展人工智能伦理审查的几种情形。

如今，对于各级监管部门以及合规主体来说，如何将人工智能伦理从道德原则框架转化为可操作、可预期、可计算的伦理合规实践，是一个亟待解决的实务问题。近来，中央有关部门正在就制定专门的人工智能伦理审查规定的议题通过多种方式征求各界意见。那么，我国该如何在“可操作、可计算、可预期”的思路下构建人工智能伦理审查制度呢？鉴于生物医学伦理[①]及其伦理审查实践较为成熟，可以将其作为重要的借鉴对象，展开相关的讨论。大致上，可以从四个方面展开论述：第一，两者具有怎样的共同点。第二，两者有什么区别，人工智能伦理具有什么特征。第三，生物医学伦理的制度实践及其理论讨论有什么教训需要吸取。第四，生物医学伦理的制度传统有哪些方面可以被人工智能伦理实践直接继承、学习。

人工智能伦理与生物医学伦理的延续性

在应用伦理学界一直存在争议的是，人工智能技术的发展是否提出了新的、独特的伦理问题，抑或人工智能伦理只是重复了在更成熟的领域，如生物医学领域遭遇的相同伦理困境？[②]对此，大部分曾经研究生物医学伦理的学者都会论证两者之间存在高度的一致性或者延续性，并认为其应统合在应用伦理学的理论体系之中。[③]

如果对生物医学伦理通说的四大原则，尊严、有利、无伤以及公正，与各国立法和政策文件以及学术界所列的人工智能伦理实体原则进行比较，确实可以得出这样的结论：人工智能伦理与生物医学伦理的

① Bioethics 在我国一般表述为生物伦理或者生命伦理。为了在中文语境里顺利展开跨国界式的学术研讨，此处暂将其翻译成“生物医学伦理”。

② Jaana Hallamaa, Taina Kalliokoski, “AI Ethics as Applied Ethics,” *Frontiers in Computer Science*, vol.4, 2022.

③ T. Beauchamp, J. Childress, “Principles of Biomedical Ethics: Marking its Fortieth Anniversary,” *The American Journal of Bioethics*, vol.19, no.11, 2019, pp.9–12.

基本原则存在较强程度的对称性和一致性。（见表 1）换言之，在法律框架内，可以采取类推等方式，从生物医学伦理审查的实体性规则中，推导出人工智能伦理审查制度的相关原则和规则。

首先，关于人类的自主性原则。正如基于人的尊严原则，个人有权就其医疗保健做出自主决定一样，人类用户也应该对他们与人工智能系统的交互有足够的控制权。这包括需要向个人提供来自人工智能 / 算法透明度的信息、个人数据收集和使用的知情同意，以及人们选择退出此类数据收集或调整人工智能系统内偏好的能力（如被遗忘权）等。

其次，关于有利和无伤原则。人工智能系统的设计和部署应最大限度地提高人类福祉并最大限度地减少伤害。这包括确保人工智能技术用于增进人类福祉，促进基本人权（如人身安全和保障），并有效应对偏见、歧视和社会差异、环境可持续性和经济差距等社会挑战。AI 开发人员和用户都有责任防止与 AI 系统相关的伤害并降低风险。这包括解决人工智能算法中的偏见和歧视，确保人工智能驱动技术的安全，以及实施保障措施以防止不良行为者意外后果或恶意使用人工智能。

再次，关于正义原则。公平和公正的伦理原则在人工智能伦理中至关重要。这涉及确保无论种族、性别、社会经济地位或地理位置 / 国籍等因素如何，人工智能技术的开发和部署方式能够促进所有人的机会平等和福利待遇优化。此外，应努力解决人工智能系统可及性方面的差异，包括“数字鸿沟”造成的差异。

近年来，在人工智能伦理原则和规则内容的讨论中，国内外出现了实质内容共识化的态势。尽管由于文化差异、特殊利益需求以及政治博弈等原因，各方提出的伦理规范内容存在不同程度的差别，但基本上涵盖以下原则：增进（人类）福祉、反偏私和公平、准确性、透明度和参与性、可解释性、保护隐私、可追责等。[①]目前，有关于人工智能伦理内容的讨论、争鸣仍然在持续，并随着人工智能技术的不断迭代，可能还会产生、凝练出新的重要伦理原则，但基本上只是针对技术的迭代在“内容周全性”方面进行完善，不大可能会突破类似生物医学伦理的“人类中心主义”原则。

人工智能伦理与生物医学伦理的差异

上文反思了过于强调人工智能伦理独特性的思

① 最新的文献综述可以参见宋华琳：《法治视野下的人工智能伦理规范建构》，《数字法治》2023 年 第 6 期；Bernd Carsten Stahl, Damian Eke, “The Ethics of ChatGPT : Exploring the Ethical Issues of an Emerging Technology,” *International Journal of Information Management*, vol.24, 2024。

表 1　生物医学伦理与人工智能伦理原则内容的对称性与延续性

<table>
<tr><th colspan="2">生物医疗伦理</th><th>《关于加强科技伦理治理的意见》</th><th>《人工智能伦理治理标准化指南》</th><th>《可信人工智能的伦理指南》</th><th colspan="2">《生成式人工智能服务管理暂行办法》</th><th>三原则</th></tr>
<tr><td colspan="2" rowspan="3">尊严
（尊重自主性）</td><td rowspan="2">尊重生命权利</td><td>合作</td><td>人类主体和监督</td><td rowspan="2">人格权
（包括隐私）</td><td rowspan="8">社会主义核心价值观</td><td rowspan="3">可信</td></tr>
<tr><td>隐私</td><td>隐私和数据治理</td></tr>
<tr><td>保持公开透明</td><td>透明</td><td>透明度</td><td>透明度</td></tr>
<tr><td rowspan="4">有利
（行善）</td><td rowspan="3">无伤
（不伤害）</td><td rowspan="3">合理控制风险</td><td>外部安全</td><td rowspan="2">技术稳健性和安全性</td><td rowspan="2">真实性、准确性、客观性、多样性</td><td rowspan="3">安全</td></tr>
<tr><td>内部安全</td></tr>
<tr><td>可问责</td><td>可问责</td><td>可靠性</td></tr>
<tr><td rowspan="2">有利
（行善）</td><td rowspan="2">增进人类福祉</td><td>可持续性</td><td>环境福祉</td><td rowspan="2">促进创新</td><td rowspan="5">负责任</td></tr>
<tr><td>以人为本</td><td>社会福祉</td></tr>
<tr><td colspan="2" rowspan="3">公正</td><td rowspan="3">坚持公平公正</td><td rowspan="2">公平</td><td>非歧视性与公平</td><td>防止歧视</td></tr>
<tr><td rowspan="2">多样性</td><td>公平竞争</td></tr>
<tr><td>共享</td><td>尊重知识产权</td></tr>
</table>

注：作者自制。

路。但是，在伦理审查的实务中，也需要警惕完全将生物医学伦理和人工智能伦理视为一体，并在人工智能伦理合规实践中简单复制生物医学伦理审查实践的保守主义思路。[①]笔者曾运用“事物本质”的概念工具，初步归纳了人工智能伦理因其技术特征而区别于生物医学伦理的三大制约性条件：道德规则的技术可嵌入性、更强的场景性以及依赖于技术过程的程序性。[②]此处沿着该文中的理论延长线继续展开讨论：

第一，合规主体上伦理责任主体单一性和“泛化”的差异。除了研发者、提供者以外，人工智能的使用者将成为伦理挑战的重点；而后者的范围可以将其理解为全世界所有的自然人和非自然人主体。不同主体对人工智能伦理的使用，一方面可以指数式提高整个社会的运转效率，但另一方面，其引起的伦理争议是纷繁复杂的，将其排除出伦理审查范围是不可想象的。换言之，和生物医学伦理相比，人工智能几乎适用于任何人类活动，这为其使用提供了无数的可能性，决定了其不可能类似生物医学伦理一样可以简单地还原为科研人员或者医生的职业伦理。它会挑战所有领域的人类业已形成的伦理道德规则，而所有的使用者都有可能成为潜在的伦理挑战者。

当然，这会引向两种讨论：一是人工智能使用阶段的伦理冲突，最终可能需要或者可以通过立法程序或者司法审查来实现，但考虑到具体应用场景的复杂性、不确定性以及效率提升的需要，可以通过各种形式的法律授权，将伦理审查的义务授权给平台型和公共管理型使用者、将伦理困境解决的大部分义务配置给自我监管主体依法成立的伦理审查委员会。二是需要区分因科技创新引发的不同类型的道德困境场景。电影《我不是药神》中的伦理争论是由仿制药是否可以与如何使用引发的，但其困境的产生根源是行政秩序维护与治病救人的价值冲突，而不是这种技术是如何影响个人和人类自主权的问题。

第二，合规内容上原则和规则侧重点的差异。以自主性原则为例，毋庸置疑，人类集体和个人的自主性是人工智能伦理原则的关键。一是个体自由意义上的自主权，包括对隐私、财产权的保护等；二是人类整体意义上的自主性，包括高度自主的人工智能系统应被设计成可以确保其目标和行为在整个操作过程中与人类价值观保持一致。尽管传统生物医学的技术研发及其应用，比如致命生物武器也可能引发人类整体被灭亡或者被其它非人类控制的危险，但人工智能系统可能会接管、威胁整个人类自主性中所体现的伦理挑战，并不是传统生物医学伦理面对的重点。

再以正义原则为例。由于人工智能系统深刻地介入并逐渐替代人类决策，它与生物医学技术本身只是决策对象、利益分配对象有着本质的区别。所以，生物医学技术的研发最多涉及分配正义，且一般是再分配阶段所考虑的内容，并非研发阶段所需思考的对象。但人工智能系统的运转本身就会消解或者异化已有的法律程序制度。比起以往应用伦理学领域处理的所有研究对象，人工智能的应用及其对社会的变革性影响更加广泛、深远。所以，每一个前沿性人工智能技术的研究、开发、部署以及使用，都有必要从人类正义或者人类现存社会秩序能否得以延续的角度展开自我审查。

第三，合规对象上“非道德性”和“有道德性”的分野。人工智能伦理与生物医疗伦理的审查对象的最大差别在于“非道德性”和“有道德性”要求。近来，国际上人工智能原则的实质内容逐渐被凝炼阐述为“可信任原则”，而“可信任”实际上也可以转述为“有道德性”的表述。

在生物医学领域，非人工智能技术意义上的生物医学技术和产品，具有鲜明的“非道德性”特征，[③]因此可以在法律合规中较顺利地应用“技术中立原则”。人工智能与传统生物医学研发和应用的区别是，它的产品和服务不只是硬件设备，而通常是极其复杂系统中的程序和应用程序，如健康和福利数据生成系统、税收系统。因此，对于人工智能技术和产品，特别是生成式

① 参见孟令宇、王迎春：《探索人工智能伦理审查新范式》，《科学与社会》2023年第4期。

② 参见李学尧：《人工智能伦理的法律性质》，《中外法学》2024年第4期。

③ Julian Savulescu, “Bioethics: Why Philosophy is Essential for Progress,” *Journal of Medical Ethics*, vol.41, no.1, 2015, pp.28-33.

人工智能系统，显然难以再简单地适用技术中立原则。

人工智能伦理审查制度构建的要点

通过上述分析，我国人工智能伦理审查制度的立法思路可以从以下方面展开。

第一，立法模式。通过前文分析，人工智能伦理的立法有必要与生物医学伦理的立法进行区别。人工智能伦理审查制度宜在科技伦理治理框架内单行立法，并在伦理审查负责主体、审查启动条件、专家构成、审查结论的法律效力和议事规则等方面做出有异于生物医学伦理的规范要求。

具体来说，就是要在现行的伦理审查体系中，在主要基于生物医学伦理审查场景的《伦理审查办法》之外，尽快起草一部专门针对人工智能伦理的平行性行政规章或者规范性文件。目前，有一种立法思路是，将《伦理审查办法》作为相关人工智能伦理审查制度的上位法。由于已有的“科技伦理”及其相关伦理审查制度（比如委员会构成）都承继了“生物医学伦理审查”的制度，为了防止后者审查思路带来的不可知的惯性思维，很有必要在短期尽快起草一部不受《伦理审查办法》限制的规范性文件或者行政规章。考虑到人工智能伦理原则的实操性关涉中国人工智能技术和产业的竞争优势，且涉及大量的程序性规范、授权性规范，因此很有必要在中短期起草一部单行性的《人工智能伦理法》，并在时机成熟时，将其作为通用型的《人工智能法》专章内容进行拟制。

第二，具体条款内容。鉴于人工智能伦理的嵌入性、场景性以及程序性特征，使得其更需要自下而上、演化型的规则生成模式，所以，在具体伦理裁断场景中，工程师与伦理专家、法学专家的有效沟通、对立性对抗、达成共识并最终转化为行为极其重要。因此，在伦理审查相关条款的设计方面，不宜将过多的条款设计资源放在实体性伦理原则的内容设计，而是应在关注程序性的视角，注重三种条款的设计：

一是关于伦理审查委员会的设置主体、职责内容和责任配置。一方面，不同于生物医学技术，人工智能技术的研发主力逐渐转变成了企业，而非高校和科研机构。因此，伦理审查委员会的主管部门应从科研机构和高校的主管部门主导，转变为由工业和信息部门主导。另一方面，从专家权力与行政权力等权力合理分配的角度，对其做更多的授权性条款设计。同时，也要配合明确人工智能伦理治理过程中的“平台型开发和应用企业”的守门人责任，并对其自我创制的相关规则进行授权性确认。[①]

二是关于伦理审查委员会的组成。为了更好地形成“对立面设置”，应对委员对选聘、构成、资格及其成员动态调整等做重点设计。其中，考虑到“有道德人工智能”的人工智能伦理审查目标及其技术可实现性，在伦理审查过程中，应注重增加人工智能技术专家的比重。人工智能伦理审查仍有“大众道德直觉”的特质，但可以通过技术途经直接落实，[②]这应该在伦理审查委员会的资格和构成中得到体现。

三是关于伦理审查委员会的议事规则。应注重评议程序、决策程序的设计。与第二点相关的是，鉴于人工智能伦理可嵌入性的技术化需求，在评议程序设计中也可以设置适合技术化的讨论流程，以防止出现伦理审查程序空转，最后沦落为“大众道德直觉限制人工智能技术进步”工具的问题。[③]

四是关于人工智能伦理审查程序启动的条件。不应将人工智能伦理审查的重点只放在研究和开发阶段、将责任主体只识别为研发者和提供者，应更多地将伦理审查责任配置给使用者，特别是决策型人工智能系统的行业使用者或者平台型使用者。这种伦理审查重点的转变，既是人工智能技术的特质所引发的，也是更好促进人工智能技术研发及其产业发展的责任优化配置的需要。

［本文系国家社科基金重点项目“支持全面创新的基础法律制度研究”（22AFX003）的阶段性成果。］

① 已有类似讨论可以参见俞思瑛等：《对话：技术创新、市场结构变化与法律发展》，《交大法学》2018年第3期。

② Adam Morton, *The Importance of Being Understood: Folk Psychology as Ethics*, London: Routledge, 2003.

③ Clara Colombatto, Stephen M. Fleming, “Folk Psychological Attributions of Consciousness to Large Language Models,” *Neuroscience of Consciousness*, vol.1, 2024.