财经法学 No. 2，2025
pp. 94 - 109

论我国人工智能领域包容审慎监管的法治维度

黄　锫*

内容提要：包容审慎监管已逐步成为我国人工智能领域的主导性监管理念。人工智能技术具有的“破坏性创新”特征决定了包容审慎监管的法治理念内核。即“包容性监管”主要是为了呵护人工智能的创新性，促进我国人工智能技术与产业的快速发展，适用于人工智能发展中遭遇的法律规则突破与法律规则空白两种情形。“审慎性监管”主要是为了应对人工智能的破坏性，防范人工智能发展中对人的权益侵害的伴生风险，适用于人工智能导致的权益侵害广度扩张与深度拓展两种情形。实现包容性监管的法治路径主要包括：通过法律解释逸脱法律规则的适用范围、灵活运用从轻/减轻/不予处罚规则、设置法定观察期等。实现审慎性监管的法治路径主要包括：将监管沙盒制度与改革试验区模式结合运用、依法划定权益保护的安全红线、设定必要的从重处罚规则等。这些研究结论都可以为我国人工智能法的立法提供参考。

关键词：人工智能立法　破坏性创新　包容审慎监管　包容性监管　审慎性监管

随着 ChatGPT、Midjourney、Sora、Dream Machine、Veo、DeepSeek 等现象级人工智能应用产品的相继问世，人工智能技术——特别是其中的人工智能大模型技术——已经成为数字经济时代创新科技的最前沿领域。然而，在推动经济社会高速发展的同时，人工智能技术的伴生风险也随之不断出现，包括但不限于侵犯训练数据中作品作者的著作权风险、〔1〕数据安全风险、〔2〕侵害个人信息权益风险、〔3〕生成虚假有害信息风险、〔4〕网络安全风险〔5〕等。由此，人工智能技术

*　黄锫，同济大学法学院教授。

本文为国家社会科学基金一般项目“人工智能大模型包容审慎监管的法治路径研究”（24BFX031）的阶段性研究成果。

〔1〕　参见焦和平：《人工智能创作中数据获取与利用的著作权风险及化解路径》，载《当代法学》2022 年第 4 期。

〔2〕　参见斜晓东：《论生成式人工智能的数据安全风险及回应型治理》，载《东方法学》2023 年第 5 期。

〔3〕　参见张凌寒：《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》，载《法律科学（西北政法大学学报）》2023 年第 3 期。

〔4〕　参见朱嘉珺：《生成式人工智能虚假有害信息规制的挑战与应对——以 ChatGPT 的应用为引》，载《比较法研究》2023 年第 5 期。

〔5〕　参见支振锋：《生成式人工智能大模型的信息内容治理》，载《政法论坛》2023 年第 4 期。

的快速演化客观上需要有效平衡促进发展和防范风险之间的关系。与此同时，包容审慎监管的理念在我国数字经济领域内迅速崛起，已经成为独具特色的中国式政府监管理念，是“监管领域对政府与市场关系的最新理论表达”〔6〕。这种监管理念自然也延伸到了人工智能监管领域，逐步成为我国人工智能领域的核心监管理念，是人工智能立法过程中不可忽视的内容。〔7〕不过，目前我国人工智能领域中包容审慎监管理念是如何出现的？它的法治理念内核包括哪些内容？实现它的具体法治途径又有哪些？这些问题都尚未明确。本文将对这些问题展开研究，以期为我国人工智能的立法提供一点理论贡献。

一、人工智能领域包容审慎监管的规范缘起

在我国，包容审慎监管的理念最初并非专门针对人工智能领域提出，而主要是为了应对移动互联网时代出现的新型数字经济业态提出。早在 2014 至 2015 年间，基于移动互联网数字平台的共享单车和网约车一经推出就引发了城市出行服务领域翻天覆地的变化。之后在日益普及的智能手机的推波助澜之下，移动互联网数字平台服务由城市出行领域不断拓展到社会生活中衣食住行的各个方面。随着这种拓展的持续深入，其形成的社会经济形态由最初的分享经济或称共享经济形态，逐步演变为平台经济形态，最终形成具有颠覆性意义的数字经济形态，并在 2020 年左右成为带动我国经济增长的核心动力。〔8〕

在这种社会经济背景之下，包容审慎监管的理念开始出现。国务院早在《2016 年推进简政放权放管结合优化服务改革工作要点》中，就已经明确提出要对新技术、新产业、新业态、新模式探索审慎监管，并要求对于一时看不准的基于“互联网＋”和分享经济的新业态应当包容发展。这一要求主要就是为了应对当时刚刚出现的共享单车和网约车等新业态，并非专门针对人工智能的发展。之后，国务院办公厅在 2017 年 1 月发布的《关于创新管理优化服务培育壮大经济发展新动能加快新旧动能接续转换的意见》中，正式提出要“探索动态包容审慎监管制度”，不过其中也并未直接提及要在人工智能领域应用这一监管理念。而在同年 7 月，国务院发布了《关于印发新一代人工智能发展规划的通知》，其中提出要建立“公开透明的人工智能监管体系”，实行“设计问责和应用监督并重的双层监管结构”，实现“对人工智能算法设计、产品开发和成果应用等的全流程监管”。从这一份关于人工智能发展的基础性政策文件来看，国务院尚未将包容审慎监管作为人工智能领域的主导性监管理念，而是更侧重于防范人工智能风险的审慎性监管，希望通过“公开透明”“设计问责”“应用监督”等措施实现对人工智能的“全流程监管”。不过，在应用人工智能技术的一些新型业态中，包容审慎监管已经初步成为主要的监管理念。除了上述共享单车和网约车的新业态之外，典型的就是自动驾驶汽车领域。如工业和信息化部在 2018 年发布的《车联网（智能网联汽车）产业发展行动计划》中，就已经明确将包容审慎作为智能网联汽车监管的主要原则。

〔6〕 刘权：《数字经济视域下包容审慎监管的法治逻辑》，载《法学研究》2022 年第 4 期，第 38 页。

〔7〕 参见宋华琳：《人工智能立法中的规制结构设计》，载《华东政法大学学报》2024 年第 5 期。

〔8〕 参见黄奇帆、朱岩、邵平：《数字经济：内涵与路径》，中信出版集团 2022 年版，第 9 页。

因此，在我国人工智能发展的初期，监管的基本理念并未被界定为包容审慎监管，后者更多地被应用于特定新型经济形态中。虽然这些新型经济形态中可能也会应用一些人工智能的技术（如网约车数字平台所使用的智能推荐算法、自动驾驶汽车领域使用的人工智能技术等），但是包容审慎监管却并未将人工智能本身作为重点适用的对象。这一点在之后国家层面发布的一些政策文件及相关规范性法律文件中也可以看出来。如 2019 年国务院发布的《关于加强和规范事中事后监管的指导意见》、国务院办公厅发布的《关于促进平台经济规范健康发展的指导意见》，以及 2020 年国务院制定的行政法规《优化营商环境条例》、2022 年全国人大常委会修订实施的《中华人民共和国科学技术进步法》中，都提及了包容审慎监管的问题，但也都没有将人工智能作为包容审慎监管的主要适用领域，最多只是将其笼统地包含在“新技术”的范畴内进行监管。

更明显的是，2022 年 3 月国家互联网信息办公室联合多部门颁布实施的《互联网信息服务算法推荐管理规定》中也未明确提及包容审慎的监管理念。算法推荐技术属于人工智能技术的重要组成部分（决策式人工智能），因此这一管理规定实际上可算是人工智能监管领域中第一部专门性的部门规章，是人工智能推荐算法技术的重要监管依据。然而，其中却并没有提及包容审慎监管的理念，更多是通过设定算法推荐服务提供者的义务来防范此类人工智能技术的风险，更偏重于审慎性监管。另一部类似的涉人工智能监管的部门规章《互联网信息服务深度合成管理规定》也存在同样的监管倾向。〔9〕

可见，虽然我国人工智能发展的顶层设计与包容审慎监管的理念大致都是在 2016—2017 年间出现，但是后者却并未被前者理所当然地采用，直到 2022 年左右两者之间都只呈现出若即若离的关联。这或许是因为，在这一时期人工智能技术并没有展现出推动经济社会发展的强大且现实的能力，只是作为数字时代类似于区块链、云计算、物联网、VR/AR、数字孪生等众多新技术中的一种而受到关注。

转折点来自 2022 年 11 月 30 日人工智能科技产品 ChatGPT 的发布。这种基于生成式人工智能技术的大语言模型应用程序几乎以一己之力颠覆了传统上将人工智能约等于人工智障的固有认知，其极为流利且准确的人机自然语言交互能力激发了将之应用于经济社会各个领域的无限想象，使人们开始真正相信“人工智能有望在人类体验的所有领域带来变革”〔10〕。之后随着迭代版本 GPT-4、GPT-4 Turbo、GPT-4o、OpenAI o1，以及文生视频大模型 Sora 的相继推出，人工智能技术已经越来越明显地处于数字时代众多新技术的最前列，成为最有可能引发新一轮科技革命、推动经济社会进入全新增长周期的创新科技，也将成为新时期国家间科技竞争的主要区域。

基于这种宏观经济社会背景，包容审慎监管理念开始真正融入人工智能领域之中。标志性的事件就是 2023 年 8 月国家互联网信息办公室联合多部委颁布实施的《生成式人工智能服务管理暂行办法》第 3 条中，将包容审慎监管正式作为人工智能大模型监管的基本理念，强调要坚持

〔9〕 这一部门规章于 2022 年 11 月 25 日颁布、2023 年 1 月 10 日正式实施，其中同样未提及包容审慎监管，也主要是通过设定深度合成服务提供者和使用者的义务来防范该项技术的风险，侧重审慎性监管。值得注意的是，这一部门规章颁布后的第 5 天，ChatGPT 才正式发布，可见其制定并未受到以 ChatGPT 为代表的人工智能大模型技术的明显影响。

〔10〕〔美〕亨利·基辛格、埃里克·施密特、丹尼尔·胡腾洛赫尔：《人工智能时代与人类未来》，胡利平、风君译，中信出版集团 2023 年版，第 16 页。

“发展和安全并重”，将“促进创新和依法治理”相结合，鼓励人工智能大模型技术的创新发展。较之于《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》这两部涉人工智能监管的部门规章偏重于审慎性监管的倾向而言，《生成式人工智能服务管理暂行办法》在设定人工智能服务提供者的义务、体现审慎性监管之外，还明确地将包容性监管也纳入人工智能领域，并专设了第二章用以鼓励与支持人工智能大模型技术的应用与发展。由此，包容审慎监管正式融入了人工智能领域的监管规范中，成为这一领域的主导性监管理念。

《生成式人工智能服务管理暂行办法》中的这一设定对于人工智能领域中包容审慎监管的发展具有标志性意义，为我国人工智能法的制定提供了重要的借鉴，产生了重要的后续影响。如我国学者在2024年拟定了两份关于人工智能立法的专家建议稿，〔11〕其中都在总则部分将包容审慎监管作为基本的监管原则。目前我国人工智能法的立法准备工作正在紧锣密鼓地推进中，可以想见，在不远将来制定的人工智能法中，包容审慎监管将会成为不可或缺的主导性监管理念。

二、人工智能领域包容审慎监管的法治理念内核

包容审慎监管之所以成为人工智能领域内的主导性监管理念，主要是因为它能够有效应对人工智能技术的“破坏性创新”（disruptive innovation）特征，正是这一特征决定了包容审慎监管的法治理念内核。

（一）“破坏性创新”决定包容审慎监管的法治理念内核

破坏性创新的概念可以追溯到20世纪40年代约瑟夫·熊彼特（Joseph A. Schumpeter）在研究产业经济变革历史时提出的“创造性毁灭”（creative destruction）这一术语，意指来自新技术、新商品、新组织形式等的竞争，使得产业发生突变的过程。这一过程不断地破坏旧的经济结构并创造出新的经济结构，最终从内部使得原有的经济结构产生革命性变化。〔12〕如果说熊彼特基于这一术语阐发的创新理论更多是对宏观经济社会结构演进的思考，那么数十年之后，克莱顿·克里斯坦森（Clayton Christensen）则是在这一思想的影响下，从微观市场竞争结构更迭的角度总结出新兴企业利用破坏性技术颠覆主流企业的创新之路，提出了破坏性创新理论。〔13〕这一理论主要是指技术推动者先通过破坏性的技术创新提供产品/服务来满足边缘市场需求，然后向主流市场逐步渗透并最终取代竞争对手的过程。〔14〕他对这一兼具毁灭与重建过程的细致分析恰好揭示了熊彼特创新理论宏大叙事下的微观市场演进形态。两者的共同点在于都意识到新技术催生新型经济社会结构的创新性，以及伴随而生的对原有经济社会结构的破坏性。

人工智能技术正是这种兼具创新性与破坏性的新技术，它能够在经济社会各个领域都产生破

〔11〕 一份是由张凌寒等学者于2024年3月提出的《人工智能法（学者建议稿）》，另一份是由周辉等学者于2024年4月提出的《人工智能示范法2.0（专家建议稿）》。

〔12〕 参见〔美〕约瑟夫·熊彼特：《资本主义、社会主义与民主》，吴良健译，商务印书馆2009年版，第146－149页。

〔13〕 参见〔美〕克莱顿·克里斯坦森：《创新者的窘境》，胡建桥译，中信出版集团2021年版，第17－31页。

〔14〕 参见斯晓夫、刘婉、巫景飞：《克里斯坦森的破坏性创新理论：本源与发展》，载《外国经济与管理》2020年第10期。

坏性创新的重大影响。在人工智能技术的赋能之下，旧有的经济社会结构会受到新兴市场主体的不断挑战，进而逐步地瓦解，同时新型的经济社会结构将会随之相继建立。例如，在人工智能大模型技术出现之后，低成本且高效产出的文本、图片、音视频等内容虽然暂时难以满足专业性的主流市场需求（如 Sora 不能满足制作完整电影的需求），但是却可以满足大量非专业性的边缘市场需求（如 Sora 可以满足制作短视频的需求）。并且随着非专业性的边缘市场需求被满足后，这一创新技术就会向专业性的主流市场需求拓展（如将 Sora 应用于完整的电影制作），在满足主流市场需求的同时彻底颠覆原有的市场经济结构（如彻底颠覆原有电影生产的市场模式）。随着人工智能技术向经济社会生活的各个领域渗透，这种破坏性创新带来的毁灭与新生过程将会无处不在，在推动经济社会高速发展的同时也不可避免地带来了巨大的伴生风险，如人工智能大模型技术的推广对训练数据中个人信息权益的侵害、〔15〕对训练数据中作品的著作权侵害〔16〕等等。人工智能技术带来的伴生风险是科技力量在市场自由竞争这只“看不见的手”〔17〕中演化的自然结果。为了在维系人工智能技术对经济社会发展的促进功能的同时，有效防范随之出现的这些伴生风险，由政府实施的包容审慎监管就成为必不可少的“看得见的手”。

包容审慎监管的内部包含了“包容”与“审慎”两个方面的内容，两者既相互区别又密切联系，是辩证统一的关系，偏废任何一方都不可取，否则会导致监管走向极端。例如有学者在研究数字经济领域的监管时就曾指出，这一领域中曾将“包容审慎监管”作为反垄断法实施的政策目标，但是却在现实中蜕变为对数字经济领域中违法行为的放任自流。〔18〕这实际上就是因为过度强调了监管的包容面，而偏废了监管的审慎面所导致的。因此，包容审慎监管应当区分为“包容性监管”和“审慎性监管”两个组成部分，两者同等重要、缺一不可。〔19〕

由此，人工智能领域的包容审慎监管也可以区分为“包容性监管”和“审慎性监管”两个组成部分，它的法治理念内核应是：包容性监管主要用来呵护人工智能的创新性，通过塑造良好的市场营商环境，为人工智能技术和产业的快速发展提供推动力，助力实现我国在人工智能领域的领先地位；审慎性监管主要用来防范人工智能的破坏性，通过划定合理的安全红线，防止人工智能的发展损害作为主体的人的最重要权益，助力实现我国人工智能的以人为本与向善发展。包容性监管与审慎性监管相辅相成，共同应对人工智能技术所具有的破坏性创新的特征，致力于实现发展人工智能与防范伴生风险之间的平衡关系。

（二）为何需要通过包容性监管呵护人工智能的创新性？

包容审慎监管中的“包容性监管”主要用来呵护人工智能的创新性。在当今世界，人工智能已经成为数字时代创新科技的最前沿，也是构建我国新质生产力最重要的技术基础。因此，呵护人工智能的创新性、促进人工智能技术与产业的快速发展便构成了我国政府监管介入这一领域的

〔15〕 参见黄锫：《生成式 AI 对个人信息保护的挑战与风险规制》，载《现代法学》2024 年第 4 期。

〔16〕 参见陶乾：《基础模型训练的著作权问题：理论澄清与规则适用》，载《政法论坛》2024 年第 5 期。

〔17〕〔英〕亚当·斯密：《国富论》（下），郭大力、王亚南译，译林出版社 2011 年版，第 24 页。

〔18〕 参见叶卫平：《数字市场反垄断法实施政策目标反思》，载《财经法学》2024 年第 6 期。

〔19〕 有学者在研究金融科技的包容审慎监管时提出了类似的区分。参见廖凡：《论金融科技的包容审慎监管》，载《中外法学》2019 年第 3 期。

基本底色，〔20〕而这也正是包容审慎监管中包容性监管的主要功能。

具体而言，包容性监管主要用来应对人工智能创新性可能会遭遇的法律规则突破与法律规则空白两种情形。

1. 法律规则的突破

人工智能技术的创新性特征意味着它可以赋能各类社会主体摆脱现有产业经济结构的束缚，拓宽社会生产力的发展空间。然而，在这个过程中，它虽然开拓了全新的产业经济发展方向，但是却同时意味着会突破支撑现有产业经济结构的法律规则体系，如布莱恩·阿瑟（Brian Arthur）所说的那样“改变制度安排的方式”〔21〕，从而出现“非法兴起”的现象。〔22〕也即，人工智能的持续发展将会伴随着相关涉人工智能行为对现有法律规则的不断突破，形式上违法的非法行为将会持续涌现。〔23〕强调此类涉人工智能行为属于“形式上”违法，是因其虽然在行为形式上违反了现有法律规则设定的义务结构，但是却代表了新科技发展的方向，潜藏着未来社会生产力的突破口，由此具有实质上的正当性。如果此时依循“严格执法”的经典法治理念，对于此类形式违法的涉人工智能行为予以惩戒，那么监管执法权力就会成为阻碍人工智能创新发展的负面力量。

因此，为了呵护人工智能的创新性以实现其对经济社会发展的促进功能，就需要适当包容此类仅具有形式违法性的涉人工智能行为。这正是包容性监管的题中之义，它能够通过包容这些仅具有形式违法性的涉人工智能行为，来呵护人工智能技术与产业的发展。

2. 法律规则的空白

同样由于人工智能所具有的创新性，它属于突破性的科技与产业领域，其拓展方向完全有可能会出现没有法律规则可以依循的情形。法律规则主要是基于人们过往行为经验制定的行为要求，对于未来可能出现的情形即使会有所预见，也无法面面俱到。并且，法律规则天然具有的保守性也要求其不能朝令夕改，必须维持相对的稳定性，这就导致不断处于流变之中的经济社会总有可能会遭遇法律规则的空白地带。人工智能作为目前发展最为迅猛的科技与产业领域，对经济社会发展具有极为强劲的推动作用，这就使得遭遇法律规则空白地带的可能性大幅度增加。如果人工智能的发展进入法律规则的空白地带，那么依循“有法可依”的经典法治理念，引发的下意识回应就会是要求及时制定相应的法律规则予以监管。然而，创新性领域的发展本身充满了不确定性，这种不确定性不但表现在出现的创新内容可能会难以预测，而且还表现在已经出现的创新内容的更迭速度可能会很快，以至于产生监管机关和创新技术的开发者“共同无知”的状况。〔24〕这意味着如果此时急于制定相应的法律规则予以监管，那么就必然会出现法律规则与创新内容之间的不匹配或相脱节。这样不仅难以对人工智能的创新予以有效监管，而且还可能会成为阻碍其

〔20〕 学者的研究也指出我国生成式人工智能的法律治理应“以发展为导向”。参见张凌寒：《生成式人工智能的法律定位与分层治理》，载《现代法学》2023 年第 4 期。

〔21〕〔美〕布莱恩·阿瑟：《技术的本质：技术是什么，它是如何进化的》，曹东溟、王健译，浙江人民出版社 2018 年版，第 213 页。

〔22〕 参见胡凌：《“非法兴起”：理解中国互联网演进的一个视角》，载《文化纵横》2016 年第 5 期。

〔23〕 有学者也曾指出，包容创新就是要容忍新业态的“非法”状态。参见张效羽：《行政法视野下互联网新业态包容审慎监管原则研究》，载《电子政务》2020 年第 8 期。

〔24〕 参见张欣：《面向产业链的治理：人工智能生成内容的技术机理与治理逻辑》，载《行政法学研究》2023 年第 6 期。

发展的负面因素。

因此，为了避免政府监管的过早介入导致监管效能与创新实践之间的脱节，包容性监管就显得尤为必要。它能够通过允许必要的法律规则留白，给予人工智能科技及产业等创新性领域相应的发展观察期，避免过早制定的法律规则不适当地约束创新科技与产业发展的自由空间。

（三）为何需要通过审慎性监管防范人工智能的破坏性？

包容审慎监管中的“审慎性监管”主要用来防范人工智能的破坏性。在通过包容性监管呵护人工智能创新性的同时，也要意识到以人为本是人工智能伦理治理的首要原则，〔25〕我们促进人工智能发展的最终目的是要为大众谋福利，让社会中每一个个体都能受益于人工智能的发展成果，能够切实感受到人工智能发展带来的幸福感。〔26〕因此，面对人工智能所具有的破坏性创新的特性，在实施包容性监管的同时，还应通过审慎性监管防范人工智能发展伴生的对人的权益侵害风险。

具体而言，审慎性监管主要用来防范人工智能破坏性导致对人的权益侵害广度扩张和深度拓展两种情形。

1. 权益侵害的广度扩张

持续提高的人工智能技术水平在展现出对经济社会巨大推动能力的同时，也蕴藏着强大的破坏能力。推动能力和破坏能力实际上是人工智能的一体两面，推动能力越大恰恰意味着破坏能力也越大。随着人工智能的不断发展，其造成权益侵害的风险将会不断增加，这首要体现在风险波及广度方面可能会远超其他领域的权益侵害风险。

例如，人工智能大模型技术已经成为目前人工智能发展的主流技术。人工智能大模型的基本技术原理是通过超强算力学习掌握超大体量训练数据中包含的词元（token）间概率分布规律，然后依据这种概率分布规律以“预测下一个词”的方式输出用户需求的信息。〔27〕在这个过程中，基于超大体量数据集进行的预训练是人工智能大模型成功的关键要素。〔28〕超大体量的训练数据常来自开发者（或数据供应商）对互联网数据的爬取，而其中包含的作品类数据覆盖面很广，以至于几乎所有互联网上能够获取的作品类数据都可能会成为人工智能大模型的训练数据。如此庞大体量的作品使得大模型的开发者要获得每项作品的著作权许可是极为困难的，〔29〕因而通常都是在没有著作权人同意的情况下使用其作品进行大模型训练。这就导致在我国目前的著作权法律制度下，开发者处理这些作品类数据时会对超大体量作品的著作权人权益产生侵害，几乎所有数字化作品的著作权人可能都是受害者，其权益侵害风险的广度将会远超其他领域。

于是，就需要通过审慎性监管来防范人工智能对人的权益侵害广度不断扩张的趋势。正如后

〔25〕 参见韩旭至：《生成式人工智能治理的逻辑更新与路径优化——以人机关系为视角》，载《行政法学研究》2023 年第 6 期。

〔26〕 参见宋华琳：《法治视野下的人工智能伦理规范建构》，载《数字法治》2023 年第 6 期。

〔27〕 关于以 ChatGPT 为代表的人工智能大模型的技术原理，参见陈锐、江奕辉：《生成式 AI 的治理研究：以 ChatGPT 为例》，载《科学学研究》2024 年第 1 期。

〔28〕 *See* Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, p. 101, available at https://arxiv.org/abs/2108.07258, last visited on Dec. 13, 2024.

〔29〕 参见丁晓东：《论人工智能促进型的数据制度》，载《中国法律评论》2023 年第 6 期。

文将会仔细分析的，它能够基于谨慎防范人工智能伴生风险的立场，通过实验型规制的方式，〔30〕将人工智能可能造成的权益侵害风险限制在一定范围内，待到探索出适当的风险控制途径后再行推广应用。

2. 权益侵害的深度拓展

人工智能蕴藏的强大破坏能力还体现在其造成的权益侵害深度可能也会远超其他领域。仍然以人工智能大模型技术为例，在进行人工智能大模型预训练的超大体量训练数据中，还包含着大量的个人信息数据。〔31〕鉴于训练数据体量对于大模型性能提高的重要性，开发者收集此类个人信息数据时就可能会出现过度收集的倾向。〔32〕当开发者将这些过度收集而来的个人信息数据用于大模型预训练之后，就会存在深度侵害信息主体隐私权的巨大风险。这是因为，人工智能大模型技术具有强大的碎片化信息整合能力，〔33〕即使大模型预训练阶段使用的个人信息数据是碎片化的，不直接涉及个体隐私信息，但在大模型采用的深度神经网络机器学习算法的强大分析能力之下，这些碎片化的个人信息数据会被深入分析、整理，隐藏在其中的个体隐私信息也会被充分整合与挖掘，甚至特定信息主体的个体隐私信息会被一览无遗地揭露。这些个体隐私信息不但会被完全暴露在大模型的开发者面前，而且大模型的终端用户也可能通过“提示词”（prompt）诱导大模型输出特定信息主体的诸多隐私信息，即大模型的通用性会造成风险向下游应用传导。〔34〕此外，最新的研究还显示，现有的技术手段已经能够从大模型中反向抽取出以 GB 为单位的数量庞大的原始训练数据，〔35〕这就导致其中包含的大量个人隐私信息也同样会被反向抽取。可见，较之传统上通常只是部分个体的部分隐私信息泄露的情形，人工智能大模型技术之下这种个体隐私信息的全方位泄露显然造成的权益侵害风险更甚。

因此，为了防范此类对人的权益侵害深度的拓展，审慎性监管就显得尤为重要。由于监管成本的固有限制，绝对地防止任何权益侵害风险的发生是不可能且没有必要的，而审慎性监管可以通过划定安全红线的监管方式，将监管资源集中在对人的最重要权益保护之上，实现安全与发展之间的平衡。

三、人工智能领域包容性监管的法治路径

如前文所述，包容性监管主要是为了应对人工智能领域出现的法律规则突破与法律规则空白两种情形，以下分别论述这两种情形中实现包容性监管的具体法治路径。

---

〔30〕 参见郭传凯：《人工智能风险规制的困境与出路》，载《法学论坛》2019 年第 6 期。

〔31〕 See Laura Weidinger et al., *Taxonomy of Risks posed by Language Models*, FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, p. 217.

〔32〕 参见林伟：《人工智能数据安全风险及应对》，载《情报杂志》2022 年第 10 期。

〔33〕 参见郭春镇：《生成式 AI 的融贯性法律治理——以生成式预训练模型（GPT）为例》，载《现代法学》2023 年第 3 期。

〔34〕 参见刘金瑞：《生成式人工智能大模型的新型风险与规制框架》，载《行政法学研究》2024 年第 2 期。

〔35〕 See Nicholas Carlini et al., *Extracting Training Data from Large Language Models*, available at https://arxiv.org/abs/2012.07805, last visited on Dec. 13, 2024.

（一）应对法律规则突破的包容性监管

法律规则的稳定性与创新科技的变动性之间总是存在着天然的张力。在相关法律规则不能及时调整与修改的情形下，我们应当通过法律规则内的包容性监管来应对涉人工智能行为对法律规则的可能突破，其中主要的方式就是通过法律解释的途径在既有法律规则框架内实现对人工智能发展的呵护。

1. 通过法律解释逸脱现有规则的适用范围

法律规则中存在着大量的不确定法律概念，〔36〕行政监管机关在依据法律规则实施监管行为时必然会涉及对这些不确定法律概念的解释，以便使监管对象的行为事实涵摄入解释后所得的法定事实构成要件之中，〔37〕并进而决定实施相应的监管措施。不确定法律概念的解释存在弹性空间，这就为行政监管机关通过法律解释使涉人工智能的行为事实逸脱〔38〕于现有法律规则的适用范围之外、实施包容性监管提供了可能性。

例如，根据《中华人民共和国著作权法》（2020 年修正，以下简称《著作权法》）第 53 条规定，侵害著作权人对作品享有的复制权并且同时损害公共利益的，行政监管机关有权责令停止侵权并对侵害人实施行政处罚。该条文设定了行政监管机关对侵害作品复制权行为实施监管的法定事实构成要件（即侵害复制权并损害公共利益）及具体的行政监管措施（即行政处罚）。依据这一法律规则，人工智能大模型的开发者在进行大模型预训练时就可能会受到行政监管机关的行政处罚。因为大模型预训练所需的超大体量数据通常是开发者通过网络爬虫技术从互联网上复制获取，或者通过购买专门数据聚合商提供的数据集获取，〔39〕其中不可避免地包含着对作品类数据的复制行为。大模型预训练数据的超大体量特性使得开发者在客观上就无法逐一获得所有著作权人的复制许可，也就至少在形式上侵害了著作权人对作品享有的复制权。并且，训练数据包含的作品数量极为庞大，几乎涵盖互联网上能够获取的所有数字化作品，涉及著作权人的范围极为广泛，因此往往很容易被认定为侵害了公共利益。然而，如果此时行政监管机关以保护公共利益的名义介入监管，依据《著作权法》的规定对大模型开发者使用作品类数据进行大模型预训练的行为予以行政处罚，那么就可能会造成寒蝉效应，导致开发者难以获取足量的训练数据进行大模型开发活动，进而阻碍我国人工智能大模型技术的发展。

面对这一监管困境，如果要通过包容性监管呵护我国人工智能大模型技术的创新性发展，增强开发者对作品类训练数据的可得性。那么，在修改《著作权法》相关规则的成本过高、不易操作的前提下，行政监管机关就可以考虑采用法律解释的方法，对相关不确定法律概念进行灵活解释，从而使大模型预训练使用作品类数据的行为事实逸脱出《著作权法》相关规则的适用范围。如有学者认为可以将人工智能大模型预训练中对作品的复制行为解释为“非作品性使用”，从而将其排除在著作权法的适用范围之外。〔40〕这一思路实质上就是通过对《著作权法》中的“复制”

〔36〕 参见〔德〕哈特穆特·毛雷尔：《行政法学总论》，高家伟译，法律出版社 2000 年版，第 133 页。

〔37〕 参见〔德〕奇佩利乌斯：《法学方法论》，金振豹译，法律出版社 2009 年版，第 131 页。

〔38〕 参见熊樟林：《论裁量基准中的逸脱条款》，载《法商研究》2019 年第 3 期。

〔39〕 参见陶乾：《基础模型训练的著作权问题：理论澄清与规则适用》，载《政法论坛》2024 年第 5 期。

〔40〕 参见刘晓春：《生成式人工智能数据训练中的“非作品性使用”及其合法性证成》，载《法学论坛》2024 年第 3 期。

这一不确定法律概念进行限缩解释，将大模型预训练时复制作品的行为事实排除在著作权法律规则的适用范围之外，证成了大模型预训练使用作品类数据的合法性，进而行政监管机关也就无需对开发者实施处罚。

2. 通过法律解释灵活适用从轻/减轻/不予处罚规则

虽然通过对不确定法律概念的解释能够在特定领域实现对涉人工智能行为的呵护，但是并非所有的情境中，行政监管机关都可以通过这种方式将涉人工智能的行为事实排除在法律规则的适用范围之外。当涉人工智能的行为事实难以被排除在规则适用范围之外时，行政监管机关还可以通过法律解释灵活运用《中华人民共和国行政处罚法》（以下简称《行政处罚法》）中的“从轻处罚”“减轻处罚”“不予处罚”等规则来实现包容性监管。

具体而言，我国的《行政处罚法》是行政处罚领域内的“基础性法律”，除非法律有专门的例外性规定，否则各类监管领域中的行政处罚行为都应适用《行政处罚法》的规定。〔41〕在该法第四章中设定了六种应当从轻/减轻处罚的适用规则、一种可以从轻/减轻处罚的适用规则、四种应当不予行政处罚的适用规则、一种可以不予处罚的适用规则。〔42〕这些行政处罚的适用规则是实现行政处罚中过罚相当原则的规范体系，〔43〕它们能够成为减免惩戒当事人力度的法定依据，并可适用于包括人工智能监管在内的各个行政监管领域。并且，这些行政处罚规则的适用条件中大量使用了诸如“消除/减轻”“危害后果”“轻微”“及时改正”等不确定法律概念。〔44〕这些不确定法律概念都存在着富有弹性的解释空间，行政监管机关可以通过对这些概念的灵活解释达到降低监管惩戒力度的目的，从而依法实现人工智能领域内的包容性监管。

我们可以用网约车的监管事例来予以说明。网约车的普及在很大程度上依赖于互联网数字平台采用的人工智能算法统筹调度，是决策式人工智能技术的经典应用之一。在网约车兴起的初期，由于接入了大量未获道路运输经营许可的私家车主，导致行政监管机关常依法对车主处以高达数万元的行政罚款。如果单从这些车主未获经营许可载客的行为形式上而言，确实属于未经许可实施客运经营的违法行为，很难通过法律解释将这一行为事实排除在相关道路运营规则的适用范围之外。虽然这种依托于人工智能算法统筹调度的新型客运经营行为确实违反了当时生效的法律规则中的禁止性规定，但是行政监管机关对相关主体施加高额罚款所产生的一般威慑（general deterrence）效应〔45〕也会在客观上阻碍人工智能技术的应用推广。

在这种情况下，行政监管机关如果要呵护人工智能技术的创新发展，实现包容性监管，那么就可以通过对行政处罚适用规则进行灵活解释以减轻甚至免除对当事人的惩戒。例如，行政监管机关可以通过解释认定网约车主停止载客的行为符合《行政处罚法》第32条规定的“主动消除或者减轻违法行为危害后果”的情形，从而依法对车主予以从轻/减轻行政处罚。或者，也可以

〔41〕 参见胡建森：《论“基础性法律”的地位及其适用——以〈行政处罚法〉为例》，载《法律适用》2023年第9期。

〔42〕 分别是第30条第2分句/第32条、第31条第3分句、第30条第1分句/第31条第1分句/第33条第1款第1分句及第2款、第33条第2分句。

〔43〕 参见刘权：《过罚相当原则的规范构造与适用》，载《中国法学》2023年第2期。

〔44〕 参见谭冰霖：《论行政法上的减轻处罚裁量基准》，载《法学评论》2016年第5期。

〔45〕 参见〔英〕罗伯特·鲍德温、马丁·凯夫、马丁·洛奇主编：《牛津规制手册》，宋华琳、李鸽、安永康、卢超译，上海三联书店2017年版，第135页。

通过解释认定网约车主停止载客的行为符合《行政处罚法》第 33 条规定的“违法行为轻微并及时改正，没有造成危害后果”的情形，进而依法对其不予行政处罚。〔46〕这样的法律解释方式可能会伴随着不同的意见和争议，但是在相关法律规则不能及时修改、涉人工智能行为的形式违法性又难以否认的情况下，通过这种方式实现包容性监管，呵护人工智能技术的发展已经是相对最优的选择。

当然，通过上述法律解释的方式实现人工智能领域的包容性监管存在一个不可或缺的前提，即行政监管机关预设的监管目的就是呵护人工智能的发展，并在这一监管目的之下对不确定法律概念进行解释。否则，由于对不确定法律概念的解释本身会存在多个不同的角度，在不同角度下的解释所实现的监管目的将会大相径庭，那么也就难以顺利实现呵护人工智能发展的包容性监管。〔47〕

（二）应对法律规则空白的包容性监管

人工智能技术赋能下的数字经济社会将会实现高速乃至超高速的发展，这也就导致进入法律规则空白之地的可能性大幅度增加。人工智能的发展一旦出现无法可依的情况，行政监管机关就面临着应当如何予以有效应对的两难问题：一方面当遭遇法律规则空白之时，往往也就意味着人工智能面临着发展演化的不确定性。行政监管机关事实上难以掌握充足的信息对其实施及时有效的监管，贸然介入反而可能成为人工智能发展的阻碍。〔48〕而另一方面，如果行政监管机关此时对人工智能不予监管，那么它又会逐步深深嵌入到经济社会结构中，一旦出现实际的危害后果，对其进行监管矫正就变得极为困难。这就是现代科技监管中的“科林格里奇困境”（Colingridge's Dilemma）。〔49〕

应该如何解决人工智能领域监管中的这一两难问题？通过设置“法定观察期”或许是一条有效的包容性监管的法治路径。〔50〕“观察期”是指当涉人工智能的新技术、新产业、新业态或新模式出现且不存在相关法律规则可以适用时，行政监管机关在一定期限内不予介入干预，或者更多使用教育、劝说、建议、提示、指导等柔性方式加以引导，给涉人工智能的新生事物留下自由发展的充分空间。〔51〕同时，行政监管机关还应积极收集相关的信息、总结相应的经验，为时机成熟之后的监管介入培育基础。

观察期的设置并非意味着行政监管机关对涉人工智能的新生事物置之不理，而是通过给予这些新生事物适当的成长期，以便为有效的人工智能监管收集充足信息。因此，观察期可以在很大程度上缓解科林格里奇困境。它一方面并不贸然采用刚性监管手段介入人工智能的发展领域，避免因信息不足而导致的乱监管；另一方面又并非对人工智能的发展听之任之，而是在时刻关注的

〔46〕 类似的讨论参见谢红星：《包容审慎理念下处罚法定原则的新发展》，载《浙江学刊》2024 年第 3 期。

〔47〕 需要指出的是，通过法律解释的方式实现包容性监管始终会面临不确定法律概念的词义弹性空间有限的问题。一旦超出了词义弹性空间的范围，那么除了及时修改相关的法律规则，在监管实践中通常就只能通过处于灰色地带的选择性行政执法来缓解刚性的法律规则与流变的社会事实之间的紧张关系。参见黄锫：《为什么选择性执法？制度动因及其规制》，载《中外法学》2021 年第 3 期。

〔48〕 参见张欣：《生成式人工智能的算法治理挑战与治理型监管》，载《现代法学》2023 年第 3 期。

〔49〕 See David Collingridge, *The Social Control of Technology*, Frances Pinter (Publishers) Ltd., 1980, pp. 17 - 18.

〔50〕 参见卢超：《包容审慎监管的行政法理与中国实践》，载《中外法学》2024 年第 1 期。

〔51〕 观察期也被称为“行政执法观察期”，关于它的形成历史参见孔繁华：《行政执法观察期的实践探索与规范进路》，载《浙江学刊》2023 年第 6 期。

过程中不断为监管的有效介入做准备。因此，观察期可以视为行政监管机关介入人工智能领域开展监管活动的前期准备阶段。

当然，如果在没有明确法律依据的情况下，行政监管机关径行在一定期限内对涉人工智能的新生事物暂缓监管，可能会引发不了解情况的第三方责难其放弃了自身的监管职责，行政监管机关可能会由此承担怠于履责的行政法律责任。出于避免此类误解的考量，即使在监管信息不充分的情况下，行政监管机关也可能会“被迫”积极地对涉人工智能的新生事物介入监管，导致过早地实施不成熟的监管措施，反而成为阻碍人工智能发展的力量。为了使行政监管机关能够不受干扰地在观察期内对遭遇法律规则空白的涉人工智能新生事物暂缓监管，就需要通过正式的立法为这种暂缓监管行为提供法定依据，使观察期成为行政监管机关暂缓监管涉人工智能新生事物的“合法期间”。因此，建议在我国人工智能立法中，明确设定人工智能监管的观察期制度，授予行政监管机关对涉人工智能新生事物暂缓监管的法定权力，给予行政监管机关实施此类包容性监管的充分法律支撑。

同时，为了避免法定观察期的制度设计沦为行政监管机关怠于履行职责的借口，还应在立法时考虑设置相应的约束性机制，督促行政监管机关在必要时积极履行职责。例如，可以考虑配套设置法定观察期内由行政监管机关实施的评估制度，要求其定期对涉人工智能新生事物的发展状况、暂缓监管的理由、恢复监管必要性与可行性等内容进行评估，制作详细的评估报告，并定期向公众主动公开。这种定期评估制度可以督促行政监管机关对涉人工智能新生事物保持持续关注，一旦出现相关风险时就能够及时介入监管，防止危害后果的出现或扩大。这样的配套制度设置也正是体现了包容性监管与下述审慎性监管之间的辩证统一。

四、人工智能领域审慎性监管的法治路径

如前文所述，审慎性监管主要是为了应对人工智能领域出现的权益侵害广度扩张和权益侵害深度拓展两种情形，以下分别论述这两种情形中实现审慎性监管的具体法治路径。

（一）应对权益侵害广度扩张的审慎性监管

在我国，审慎性监管中有效应对人工智能领域权益侵害广度扩张的主要法治途径是将舶来的监管沙盒制度（regulatory sandbox）与本土的改革试验区模式进行结合运用。

监管沙盒制度源于2015年英国金融行为监管局（FCA）为应对金融科技发展而实施的监管创新〔52〕（欧盟的《人工智能法案》已将监管沙盒制度正式纳入人工智能监管体系）。在这种监管制度下，英国的行政监管机关依据法律授权允许符合特定条件的金融科技企业进入监管沙盒，在有限的业务牌照下利用真实或模拟的市场环境开展金融创新业务，经过一段时期的测试证明有效后再全面推广。〔53〕这种创新监管制度与我国经济社会发展中常采用的改革试验区模式具有异曲同工之处。改革试验区模式是指我国为了深化经济社会领域的改革，在特定行政区域内通过试点

〔52〕 参见张红：《监管沙盒及与我国行政法体系的兼容》，载《浙江学刊》2018年第1期。

〔53〕 参见许多奇：《金融科技的“破坏性创新”本质与监管科技新思路》，载《东方法学》2018年第2期。

来探索某些领域改革开放的路径或可能性，以在积累相关经验的基础上，为其他区域的发展树立样板，并将其试点经验推广到其他区域，甚至据此推动新制度在全国的确立。舶来的监管沙盒制度与本土的改革试验区模式的共同点在于都采用了试点成功后全面推广的思路，不同点则在于前者通常是由行政监管机关在某个产业领域（如金融领域）中选取特定的几家企业赋予资格后开展试点，后者则通常由国家选取特定行政区域并给予灵活适用法律、法规或规章的权力，通过制定试验性立法〔54〕开展多领域的改革试点后再行推广，覆盖面要远远广于前者。

为了防范人工智能的发展导致对人的权益侵害范围的扩张风险，我们可以将这两者相结合，在改革试验区内通过涉人工智能的试验性立法设定监管沙盒制度，从而实现审慎性监管的法治化。试验性立法是改革试验区内贯彻实现“改革只能在法律的范围内进行”〔55〕的法治要求的主要途径。全国人民代表大会及其常委会通过发布“授权决定”的方式赋予特定机关在改革试验区内通过试验性立法变通法律、行政法规、部门规章的权力，为改革试点提供合法性依据。例如，全国人民代表大会常务委员会曾于 2021 年 6 月作出授权决定，授权上海市人民代表大会及其常委会制定浦东新区法规，可以对法律、行政法规、部门规章作出变通规定（《中华人民共和国立法法》2023 年修订时已将其纳入正式立法）。由此，浦东新区作为改革试验区，就能够通过试验性立法对法律、行政法规、部门规章中设定的内容进行变通实施。我们可以运用这一授权，通过浦东新区法规在浦东新区改革试验区内建立人工智能的监管沙盒制度，允许符合条件的人工智能企业进入监管沙盒，在受控的前提下开展人工智能创新技术和产品的研发。这样即使人工智能技术或产品存在侵害人的权益的风险，也能被有效地限定在监管沙盒所涉企业以及改革试验区域范围内。

以人工智能大模型的开发为例，其开发所使用的超大体量训练数据中包含了大量的作品类数据，存在着侵害著作权人权益的范围不断扩张的风险。对此，我们可以将监管沙盒制度与改革试验区相结合，通过制定浦东新区法规在浦东新区改革试验区内设定人工智能大模型的监管沙盒制度，将符合条件的大模型开发企业纳入监管沙盒，在行政监管机关严格监督的前提下开展大模型的预训练活动。进入监管沙盒的企业数量有限且经过严格删选，因此行政监管机关就能具备充足的行政资源监管企业利用作品类数据进行大模型预训练的活动，进而也就有能力将可能出现的侵害著作权风险限制在一定范围之内，防止其权益侵害范围的扩张。简言之，这种“试验型规制制度是降低实施中错误成本的有效方法”〔56〕。

（二）应对权益侵害深度拓展的审慎性监管

人工智能技术具有的强大数据分析与挖掘能力潜藏着对人的权益的深度侵害风险，我们应当通过以国家强制力作为后盾的行政监管机关来防范此类权益侵害深度的拓展。

1. 依法划定权益保护的安全红线

审慎性监管并非绝对禁止一切可能产生权益侵害的人工智能，否则会彻底禁锢人工智能的发

〔54〕 参见黎娟：《“试验性立法”的理论建构与实证分析——以我国〈立法法〉第 13 条为中心》，载《政治与法律》2017 年第 7 期。

〔55〕 李龙：《中国特色社会主义法治体系的理论基础、指导思想和基本构成》，载《中国法学》2015 年第 5 期，第 20 页。

〔56〕 靳文辉：《试验型规制制度的理论解释与规范适用》，载《现代法学》2021 年第 3 期，第 129 页。

展，而是要在谨慎权衡利弊的基础上，依法划定人工智能开发活动不可以突破的安全红线，〔57〕聚焦于保障最为重要的权益类型，使人工智能对于经济社会发展所产生的总体效益高于伴生的权益侵害损失，从而实现卡尔多-希克斯标准（Kaldor-Hicks criterion）意义上的效率。〔58〕

以我国人工智能大模型技术的发展为例，就其技术特性而言，需要在大模型预训练阶段使用超大规模的训练数据才能达到理想的机器学习效果，而目前互联网上可以获取的训练数据中包含了大量个人信息数据。如果为了保护信息主体的个人信息权益，严格禁止在大模型预训练中使用这些个人信息数据，那么就可能在很大程度上减少训练数据的体量，进而影响大模型预训练的学习效果。但是如果完全放任这些个人信息数据被用于大模型预训练，又可能会对信息主体的个人信息权益造成侵害。并且，在大模型强大的数据分析能力之下，这种权益侵害的深度会不断拓展甚至失控。因此，我国的审慎性监管需要在法律框架内，谨慎考虑安全与发展之间的平衡关系，实现对用于大模型训练的个人信息数据“从权利保护到公平使用”〔59〕的规制目标。

由此，基本的审慎性监管思路是将个人信息数据所涉及的权益分为人格性权益和财产性权益两个部分，并将保护的重点置于人格性权益之上，同时适当放宽对其中的财产性权益的保护。这是因为，个人信息中的人格性权益涉及信息主体作为人的基本尊严，难以用经济效益加以衡量。而个人信息的财产性权益对于信息主体而言所占权重通常并不大，且人工智能技术发展所带来的总体经济社会增益很高，从整体上而言足以补偿相应的个人信息财产性损失，满足卡尔多-希克斯效率标准的要求。因此，在将个人信息数据用于人工智能大模型预训练时，权益保护的安全红线就可以划定在保护个人信息中的人格性权益之上。也即，审慎性监管的重点在于保障大模型训练数据中信息主体的人格性权益，只要不是损害信息主体人格性权益的大模型数据处理行为都应属于允许的范围。

依据这一安全红线的划定思路，结合前文所述的监管沙盒与改革试验区制度，我们可以在改革试验区内通过试验性立法对《中华人民共和国个人信息保护法》进行变通规定，允许纳入监管沙盒的企业在不侵害信息主体人格性权益的前提下，自由利用个人信息数据进行人工智能大模型的开发。这样一方面为人工智能企业开发大模型松开了法律规则上的束缚，能够更充分地挖掘个人信息数据中蕴藏的价值；另一方面也能守住保护信息主体人格性权益的安全红线，在监管沙盒和改革试验区的制度辅助之下将可能产生的损害风险限制在最小范围内。

2. 设定必要的从重处罚规则

权益保护安全红线的划定虽然重要，但仅仅只是设定了人工智能领域应重点保护的法律权利和相关主体不得违反的法律义务。如果没有相应的法律责任机制予以保障，那么这种法律权利和法律义务的设定就可能会沦为纸面上的法，而无法成为实际被遵守的“活着的法”〔60〕。因此，

〔57〕 参见刘乃梁：《包容审慎原则的竞争要义——以网约车监管为例》，载《法学评论》2019年第5期。

〔58〕 卡尔多-希克斯效率是指受益者的收益能够潜在地（无需实际支付）补偿受损者的损失，也被称为潜在意义上的帕累托优势效率。参见〔美〕理查德·A·波斯纳：《法律的经济分析》（上），蒋兆康译，中国大百科全书出版社1997年版，第16页。

〔59〕 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》，载《比较法研究》2024年第4期，第94页。

〔60〕〔德〕托马斯·莱赛尔：《法社会学导论》（第5版），高旭军等译，上海人民出版社2011年版，第68页。

在依法划定人工智能发展中安全红线的同时，审慎性监管还应当通过必要的法律责任机制确保法律权利能够得到切实保护，以及法律义务能够得到切实履行，其中尤为重要的法治路径之一就是针对突破权益保护安全红线的行为依法设定从重处罚规则。

从重处罚是指在法定行政处罚裁量幅度内靠近上限实施处罚。正如在包容性监管中可以通过从轻/减轻/不予处罚规则来呵护人工智能的创新性发展一样，在审慎性监管中同样可以通过从重处罚规则来应对涉人工智能行为突破安全红线的风险。也即，我们可以将人工智能领域中突破安全红线的行为设定为行政处罚的“从重情节”，〔61〕予以从重行政处罚。之所以需要设定从重处罚规则，主要因为：其一，在人工智能技术的破坏性创新特征的加持之下，涉人工智能行为突破权益保护安全红线的深度会持续拓展且难以预测底部。因此，需要通过从重处罚规则来应对这种权益侵害深度的不确定性，以便在人工智能技术或产业发展过程中出现不可预测的重大权益侵害时，依法实施与危害程度相适应的行政处罚，使责任主体承担过罚相当的法律责任。〔62〕其二，专门针对突破权益保护安全红线的涉人工智能行为设定从重处罚规则，实际上也是在向人工智能技术与产业的相关主体传递一种“信号”，〔63〕即行政监管机关对这些权益保护安全红线的重视程度以及牢牢守住权益保护安全红线的重要性，提醒这些主体突破这一红线所要承担的严格法律责任，从而能够在事先尽最大可能地预防发生突破权益保护安全红线的行为。

在我国目前的法律体系中，虽然已有部分单行法律设定了从重处罚规则，但在《行政处罚法》中，除了突发事件的特殊情形之外（第 49 条），并未设定从重处罚的一般性规则。这意味着在目前人工智能法缺位的情况下，对于突破权益保护安全红线的涉人工智能行为，行政监管机关无法依据《行政处罚法》的一般性规定在必要时实施从重处罚，这就可能会导致对此类涉人工智能的违法行为难以实施过罚相当的行政处罚。因此，面对人工智能破坏性创新所带来的权益侵害深度拓展的不确定性，如果不能通过修改现行《行政处罚法》增加从重处罚的一般性规则，那么在未来制定的人工智能法中针对突破权益保护安全红线的涉人工智能行为，设定特定条件下的从重处罚规则或许是最为可行的方案。

五、结　语

综上所述，虽然包容审慎监管最初并非专门针对人工智能领域提出，但是随着大模型技术出现之后迎来的人工智能爆发式增长，包容审慎监管将会成为我国人工智能领域的主导性监管理念。人工智能技术所具有的破坏性创新特征决定了包容审慎监管的法治理念内核，使其区分为包容性监管与审慎性监管两个部分。其中包容性监管的主要目的在于呵护人工智能的创新性，以便促进人工智能的快速发展，打造我国在新经济周期中科技与产业的领先地位。包容性监管主要适用于人工智能发展可能遭遇的法律规则突破与法律规则空白两种情形，我们可以通过法律解释逸脱现有法律规则的适用范围、灵活运用从轻/减轻/不予处罚规则、设置法定观察期等法治途径予

〔61〕 参见张淑芳：《行政处罚应当设置“从重情节”》，载《法学》2018 年第 4 期。

〔62〕 参见金成波：《从重处罚设立的必要性及其制度构造》，载《行政法学研究》2022 年第 4 期。

〔63〕 参见〔美〕埃里克·A·波斯纳：《法律与社会规范》，沈明译，中国政法大学出版社 2004 年版，第 48－49 页。

以实现。审慎性监管的主要目的在于应对人工智能的破坏性，以便防范人工智能的伴生风险，平衡我国人工智能发展与安全之间的关系。审慎性监管主要适用于人工智能的发展可能导致对人的权益侵害广度扩张与深度拓展两种情形，我们可以通过将舶来的监管沙盒制度与本土的改革试验区模式结合运用、依法划定权益保护的安全红线、设定必要的从重处罚规则等法治途径予以实现。这些关于人工智能领域包容审慎监管的具体法治途径，可以为我国人工智能法的立法活动提供参考，从而建构一个既能有力促进我国人工智能技术与产业发展，又能有效防范人工智能伴生风险，且具有中国特色的人工智能监管法律体系。

Abstract: Inclusive and prudent regulation (IPR) has gradually become the dominant regulatory concept in the field of artificial intelligence (AI) in China. The "disruptive innovation" feature of AI technology determines the core concept of IPR. That is, "inclusive regulation" is mainly to protect the innovation of AI and promote the rapid development of AI technology and industry in China. It is applicable to the two situations where AI breaks through legal rules and where there are legal rule blanks. "Prudent regulation" is mainly to deal with the destructiveness of AI and prevent the concomitant risks of infringement on the rights and interests of human beings as the subjects in the development of AI. It is applicable to the two situations where the breadth and depth of rights and interests infringement caused by AI expand. The legal paths to achieve inclusive regulation mainly include: evading the application scope of existing legal rules through legal interpretation, flexibly applying the rules of "lighter/mitigated punishment" or "no punishment", and setting up a "statutory observation period", etc. The legal paths to achieve prudent regulation mainly include: combining the regulatory sandbox system with the model of reform pilot areas, legally delimiting the safety red line for rights and interests protection, and setting necessary rules of heavier punishment, etc. All these can provide references for the legislative activities of the "AI Law" in China.

Key Words: artificial intelligence act, disruptive innovation, inclusive and prudent regulation, inclusive regulation, prudent regulation

(责任编辑：林洹民)