大模型智能体的行为边界与规制

□ 杜心蕾

从聊天机器人走向能够自主规划和执行任务的智能体，是大模型应用的重要发展方向，与传统软件按照预设程序执行任务不同，智能体能够理解自然语言指令，将任务拆分为多个步骤，自主选择工具和路径，并与外部网站、应用和数据接口发生交互。这意味着，人机关系正在由单纯的工具使用关系转向具有一定代理特征的任务执行关系。由此产生的治理问题也不再局限于模型生成内容是否安全，而是涉及智能体对外行动的权限来源、行为边界及责任归属。其中的一项基础问题就是用户的一次概括性指令，究竟能够授权智能体实施哪些具体行为以及他的行为边界到底在哪。

传统互联网服务中，平台主要提供信息。智能体则改变了这一过程，用户可能只提出“比较几种商品价格并完成购买”这样的目标性指令，至于访问哪些网站、调用哪些接口、是否登录账号、是否接受第三方协议以及向何方提交信息，都可能由智能体自行决定。用户指令具有概括性，而智能体执行却包含一系列具体行为。目前，服务提供者往往通过用户协议取得概括授权，允许智能体为完成任务而访问必要网站、调用必要工具或者处理必要数据。但是，概括授权却不能无限延伸至任务执行过程中的所有行为。尤其是支付、缔约、提交个人数据、使用身份凭据等行为，直接涉及用户财产、隐私和交易责任，不能仅凭用户协议中的抽象条款，就推定用户已经授权智能体实施。

因此，需要预先确立授权的合理边界，以普通用户对于完成该任务所需合理步骤为基础下达用户指令。例如，用户要求查询某本图书的评分，可以合理预见智能体访问相关图书页面，却难以据此认定其授权智能体抓取整站、绕过登录限制或者大规模下载数据。凡明显超出任务需要的操作，不应纳入授权范围。同时，应区分一般操作与重大权益操作。比如，查询公开信息、比较价格等低风险行为，可以在授权范围内自动完成。登录账号、向第三方披露个人信息、签署协议、购买商品、支付款项等行为，则应提高授权门槛，通过醒目提示、二次确认或者单独授权，使用户在关键节点保有决定权。这样既能够维持智能体自动执行任务的技术优势，也能够防止自动化便利以削弱用户控制为代价。

技术能力不等于行为授权。智能体授权并非只是用户与机器之间简单的两方关系，其运行至少涉及三个层次：用户向服务提供者发出任务指令，服务提供者为智能体配置工具和权限，智能体再向第三方网站、应用或者接口发出请求。三层权限并非天然一致，用户可能只希望完成一次查询，但是服务提供者却为智能体配置了更广泛的数据访问能力，并且在用户允许读取某项信息后，智能体可能会自主选择新的目标和工具。由此需要警惕一个基本误区：技术上“能够做到”，并不意味着法律意义上“有权做到”。

智能体的权限配置应坚持最小必要原则。服务提供者不能因为技术上能够开放抓取、登录、支付、跨站调用等能力，就默认将这些能力全部配置给智能体。完成信息查询不需要默认取得支付权限，比较商品价格不意味着可以自动接受第三方协议，读取用户日历也不意味着可以向无关平台传输其中的数据。将权限控制前置到产品架构和系统设计之中，比发生争议后再追究责任更有效。

还应做好区分用户主动请求与服务提供者主动抓取工作。用户主动请求是智能体为完成用户明确提出的任务而实施的必要访问，而服务提供者主动抓取则可能是平台为了训练模型、建立索引、丰富搜索结果等自身业务开展的大规模自动化活动。二者的目的、范围和受益主体均不相同。平台不能借用户的一次具体请求，为自身的数据获取行为寻找授权依据，更不能将用户授权无限延伸至后台业务活动。平台为自身业务目的实施的行为，应当具有独立、明确的合法基础。

此外，智能体越具有自主选择目标、工具和执行时间点的能力，就越需要明确的权限约束。对于可能显著增加用户风险的自主操作，应通过权限分级、风险提示和授权确认等方式，使技术自主性始终围绕用户意愿。

清晰的授权规则并非限制创新。智能体要进入购物、出行、办公、金融服务等场景，既需要用户信赖其任务执行过程，也需要第三方平台能够判断自动访问是否具有正当授权。稳定、透明、可预期的权限规则，反而能够降低各方摩擦，为智能体跨平台协作提供可信基础。

面对复杂的自动执行过程，用户权益保障不能只依赖服务协议，而应当转化为用户能够直接感知和使用的产品功能。授权机制需要嵌入产品设计，可以围绕知情、控制和救济三个环节，建立贯穿任务全过程的授权机制。

提高智能体行为的透明度。现实中，一些智能服务只展示最终结果，用户却无法了解智能体访问了哪些系统、调用了哪些接口、向第三方提交了什么信息。服务提供者应保存必要的操作日志，并以普通用户能够理解的方式提供任务记录。涉及登录、支付、提交个人数据、接受第三方协议等关键操作时，应给予实时提示，而不能将知情义务简单转化为事后查阅复杂技术日志。

保障授权状态的可控性。用户应能够暂停任务、撤回授权，并对智能体的能力设置具体边界，例如禁止访问特定类别网站、禁止代替用户接受协议、禁止调用支付工具，或者限制使用身份凭据进行跨站操作。可以设置统一的权限管理中心，将智能体能够访问的数据、工具和外部服务集中呈现，使用户能够持续查看并调整权限。对于自动续期、连续抓取、批量操作等持续性行为，则应设置更严格的确认机制。

增强执行过程的可追溯性。智能体发生越权操作后，如果用户无法了解执行路径，也无法取得相关日志，责任就容易在用户、模型服务商和第三方之间相互推诿。因此，操作记录既是保障用户知情的工具，也是认定责任的重要依据。服务提供者提供模型行为、技术架构和后台日志，对于系统运行具有更强控制能力，不能简单以“模型自主生成”或者“用户已经授权”为由转嫁风险。对于超越用户明确授权且服务提供者未尽合理限制义务的行为，应根据各方实际控制能力和过错合理分配责任。因此，授权机制不应成为纠纷发生后的补救措施，而应成为智能体产品的基础能力。要进一步形成统一的权限分级和交互规范：低风险操作在合理范围内自动执行，高风险操作触发明确确认；授权范围随任务结束而相应收缩，临时取得的数据和权限不应长期保留；智能体调用外部工具时留下可验证记录，使用户、平台和第三方能够识别行为来源和权限状态。

此外，大模型智能体正在推动人工智能由信息生成工具转向能够自主规划、调用工具和执行任务的应用形态，对大模型智能体的治理也不能只关注模型本身的安全，还应关注模型与用户、平台、第三方连接之后形成的完整行为链条，以清晰授权建立智能体可信应用基础。尤其在建立智能体授权规则的同时，需要在便利与安全、创新与规范之间形成平衡。不能要求每一个技术步骤都由用户逐一确认，否则会削弱智能体的自动化价值；也不能以一次概括同意覆盖所有后续行为，使用户逐渐失去对数据、财产和身份的有效控制。较为可行的路径，是以合理可预见范围划定一般授权边界，以最小必要原则约束技术权限配置，以重大权益事项再次确认为风险控制机制，并通过日志和权限管理保证行为可追溯、授权可撤回。

长远来看，智能体产业的发展不仅取决于技术能力，也取决于用户能否信赖其任务执行机制。只有让用户知道智能体能够做什么、正在做什么，并能够决定其权限范围，在发生问题时又能够追溯行为和明确责任，才能为智能体进入更广泛的经济社会场景建立稳定预期。以清晰授权守住智能体对外行动的边界，以有效控制释放智能体自动执行的价值，既是保障用户权益的需要，也是推动大模型智能体规范、健康发展的重要条件。

数字观察

人工智能参与决策要坚持分级分类治理

□ 冀翠萍

人工智能进入决策领域，为推动经济社会发展带来了新机遇。决策中依托大数据分析突破了传统决策中的信息处理限制，提高了决策速度与资源配置效率，减少了人为干预偏差。但同时也带来了新的安全问题，决策失真、程序不公与决策错误规模化扩散引发的公共安全和社会信任风险正在显现。这就要求我们应厘清智能时代人工智能安全边界，建立与新的安全逻辑相适应的治理体系，从而确保人工智能安全有序地参与决策。

当人工智能开始参与决策，人类决策权部分让渡给机器后，安全逻辑便发生了深刻变化。第一，安全的焦点从系统是否被攻破，转向决策是否正当；第二，风险的形态从可见的、局部的、一次性的技术故障，转向了隐性的、可复制甚至可规模化扩散的决策错误；第三，危害的对象从财产、经济利益与便利性，转向了公平、权利与社会信任。智能时代不仅要求技术可靠，更要求决策可信，即决策权要始终处于能明确主体的状态下，决策依据能够合理说明，决策过程能够透明可监督，决策结果正确可靠，并在发生错误时能够实现责任追溯和权利救济。智能时代安全的标准，由是否按规则运行，转向决策是否值得信任、是否可掌控。因此，保障智能决策安全应坚持分级分类原则，依据人工智能参与决策的深度和领域，匹配差异化的安全要求与治理准则。

按照参与深度进行分级治理，明确智能决策的权限边界。依据人工智能在决策过程中呈现的是辅助、协同还是自主的状态，将参与决策的深度分为三个层级。第一层是辅助决策，人工智能主要提供信息、预测和建议，由人类作出最终判断并承担责任，对此应积极推进，释放其提升效率、优化决策的价值。第二层是协同决策，人类决策者与智能系统交互协作、共同作用，人工智能形成方案建议，须经人类确认方可生效，对此应加强透明审查、人工监督和结果复核，防止机器判断未经检验直接生效。第三层是自主决策，人工智能在一定范围内直接执行，人类保留监督、暂停和最终决定权，对此应审慎推进，由具名主体对决策结果承担责任。基本逻辑是参与越深、影响越重、不可逆性越强，人类的决策权重就越高、责任配置就越明确。当人机决策冲突或不一致时，机器决策应该让位于人类决策，保持人类的决策权与主体性。

按照应用领域进行分类治理，匹配差异化的治理准则。在一般性信息处理、商业运营优化等领域，人工智能主要用于辅助分析和效率提升，对个人权益和公共利益影响较小，治理重点应放在技术稳定性和规范应用上，明确负面清单，鼓励行业自律。在医疗辅助、金融风险评估、人力资源管理等领域，人工智能的判断结果直接影响个体权益，需加强对数据质量、模型公平性、结果可解释性的技术性审查，开展上线前的算法安全评估和伦理审查，保留人工复核与异议申诉渠道。在行政管理、司法辅助、社会救助等公共领域，人工智能参与的是涉及公共权力运行、公共资源配置和重大社会利益的决策，应定位为辅助工具，最终决定是由具体的责任主体作出，决策依据和过程全程留痕，以保障公共权力的正当性与公信力。基本逻辑是越涉及公平公正等价值性领域，对人工智能应用的审查就应当越审慎严格。

推动智能决策安全有序，应以分级分类为主线健全智能决策安全的治理体系，推动分级分类原则贯穿治理体系全过程，使不同参与深度、不同应用领域的智能决策，对应不同的治理机制，使决策权限有边界、决策过程可核查、决策风险能预警、错误损害可救济。

健全法律法规体系，确立决策权限的法定边界。加快推进人工智能领域专门立法，重点把散见于各部门规章中的监管规则整合上升为法律制度。一是明确人工智能风险分级分类的法定标准，综合考虑技术自主性程度、应用场域敏感程度、潜在危害的规模和不可逆程度等因素，形成动态可调的分级分类目录。二是依法界定不同参与深度、应用领域下的人工智能决策权限，明确哪些决策可以由算法辅助、哪些必须由人作出最终决定。三是确立控制程度与责任范围相匹配的原则，明晰数据提供者、模型开发者、系统部署者、具体使用者在决策链条上的责任边界，使决策问责有落点，让各类主体对自身义务形成稳定预期。

完善技术监测体系，实现决策过程全程可核查。监管对象的技术属性决定了监管手段必须技术化。与一般系统安全监测不同，智能决策监测的核心在于决策过程本身。在输入端，要监测训练数据和实时数据的质量与偏差，防止数据缺陷转化为决策失真；在模型端，定期开展算法审计和模型评测，重点检验模型偏差、鲁棒性和决策依据的可说明性，防止“评估一次、终身有效”；在输出端，对自主决策等高参与深度的系统实行持续监测，对决策结果异常集聚、明显偏离预期的情况及时核查，确保任一决策可回溯至具体环节与责任主体。同时整合算法备案系统、生成式人工智能服务备案等既有监管数据资源，形成风险线索的汇聚、比对和分办机制，改变事后被动处置的局面。

建立风险预警体系，防范决策错误规模化扩散。智能决策风险的特殊性在于，个别误判可能借助算法的统一适用被迅速复制，演化为系统性风险，预警体系必须对此作出针对性安排。应按照参与深度和应用领域设定差异化的预警阈值和报告义务，对自主决策、公共领域等高风险系统，实行强制性的安全事件报告制度，参照网络安全领域重大事件限时报告的制度实践，明确报告主体、报告时限和报告内容，隐瞒不报、迟报漏报的依法追究责任；对辅助决策、一般领域等中低风险系统，建立常态化的风险信息收集机制，依托举报受理、行业监测和用户申诉渠道汇集线索。健全跨部门、跨区域的风险研判会商机制，社会保障、医疗救助、交通调度等系统一旦出现大规模误判和错误决策，立即启动快速核查和纠偏，把扩散风险阻断在萌芽状态。

构建应急响应体系，确保决策错误可纠正可救济。制定人工智能决策安全事件应急预案，按照事件危害程度实行分级响应，明确处置流程、责任主体和协同机制。构建应急响应措施应对决策失灵的不同对策：对存在重大缺陷的决策应用，责令暂停服务、限期整改，相关业务转为人工办理；对已经作出的错误决策，组织复核纠正，涉及公民权利义务的依法启动救济程序；对可能危及国家安全和社会稳定的，依法采取下架产品、关停系统等措施。应急处置应当规范启动条件、实施程序和期限，既要防止反应迟缓贻误时机，也要防止过度处置影响正常服务。应急处置与责任追究相衔接，对滥用恶用人工智能决策的依法追责，把“确保人工智能始终处于人类控制之下”落到实处。

数字治理

完善适应人工智能技术发展促进就业的措施

□ 于 倩

人工智能作为引领新一轮科技革命和产业变革的重要驱动力，正深刻重塑着人类生产生活方式，对就业格局产生深远而复杂的影响。2026年政府工作报告提出，“完善适应人工智能技术发展促进就业创业的措施”。2026年10月，国务院办公厅印发的《关于适应人工智能技术发展助力构建就业友好型发展方式的意见》强调，积极增强人工智能创造就业机会、赋能传统岗位、提升就业质量、激发创业活力的正面效应，有效减缓人工智能应用就业替代影响，努力实现技术进步、产业升级与就业促进、民生改善相协调的包容性发展。“十五五”规划建议明确提出，全面实施“人工智能+”行动，促进高质量充分就业。这些都为应对技术变革与做好民生保障指明了方向。

人工智能对经济社会产生重要影响，新技术的发展激发了进一步创造新岗位、改造传统岗位、提升劳动价值、激活创业空间的潜力。人工智能产业本身涵盖了一个包括基础层、技术层和应用层在内的产业链条，这一产业链催生出一些包括数据标注、模型优化、传感器研发、交通系统运维等一系列新职业、新岗位就业需求。激励更多劳动者找到与自身优势更契合的岗位，为劳动者提供了更多元的职业发展通道。“人工智能+”行动持续培育新产业新业态，激发新需求，围绕垂类模型开发、数据治理、模型运维、提示工程、算法合规等方面拓展了岗位需求，催生出新兴就业岗位，为就业市场持续注入新动能。

人工智能更多是对重复性、流程性工作任务形成替代，而非对职业整体的简单取代，合理的人机协同模式，能够把劳动者从机械繁琐的事务中释放出来，转向复杂研判、创意设计、人际沟通、质量管控等高价值环节。人工智能通过提供更好的工具和知识信息，将传统岗位工作流程更迅速地呈现，促进了生产力发展和工作效率提高。例如，在医疗行业，医疗辅助机器人将部分低附加值的任务替代，使职业的核心价值得以提升，推动了行业就业结构的优化升级。人工智能应用使得就业结构向两端化延伸演进，就业重心向高附加值服务业转移，推动劳动力流向需要复杂交互、创新思维及情感投入的生产性服务业和高品质生活服务业，尤其对能够驾驭复杂系统的高技能人才需求增加，使劳动者素质不断提升。

客观来看，人工智能技术的发展，改善了作业环境并降低了劳动强度，重塑了人机协作模式，通过智能机器人将人从高温、高压、有毒或高重复性的“危繁脏重”场景中解放出来。利用人工智能自动处理数据录入、文档检索等机械性事务，将劳动者从低价值的枯燥流程中解放出来，使其能专注于更具情感交互、复杂决策和创新性的高价值工作，优化了工作内容并释放出积极创造力。人工智能辅助工具降低了复杂任务的操作门槛，打破了办公的时空限制，支持更加弹性、自主的工作模式。同时，人工智能借助大模型、智能体等工具，降低了技术、资金和人力投入成本，推动产业发展从依赖规模投入向轻量化、低成本、高效率转变，为个体创业和小微企业发展创造了更多机会。人工智能加速技术成果转化和市场需求匹配，促进创意、技术与市场有效衔接，推动更多创新想法转化为创业项目和市场价值。此外，人工智能催生新技术、新产品、新业态和新商业模式，推动产业链各环节创新重组，拓展数字内容、智能制造、专业服务等领域的创业空间。

然而，人工智能对就业的影响是一把双刃剑，在赋能就业提质的同时，也带来了新的挑战。比如，劳动力技能更新的速度与技术迭代速度存在一定差距、短期内“机器换人”的“替代效应”产生等。必须强化系统观念、注重多措并举，在发展导向、人才培养、制度保障上精准施策，有效减缓人工智能应用就业替代影响，努力实现技术进步、产业升级与就业促进、民生改善相协调的包容性发展，为广大劳动者高质量就业保驾护航。

构建就业友好型发展方式，推动人工智能发展与就业提质深度耦合、同向发力。坚持就业优先导向，强化财政、货币、产业政策与就业政策的协同性。完善财税、金融优惠政策，加大对就业友好型人工智能应用企业的支持力度。健全人工智能发展评估、就业影响监测与政策评估机制，加快构建覆盖岗位变动、技能供需、就业质量的高频监测体系，分行业、分区域设置预警阈值与政策触发规则。在重大政策制定、重大项目落地、重大生产力布局中，同步开展岗位创造、失业风险与就业质量影响评估。设立人工智能咨询专家机构，系统研判各领域技术发展趋势与潜在就业效应，将“就业友好型”指标嵌入产业政策与财政资金评价环节。加大对赋能型、互补型、创造型人工智能技术支持力度，前瞻性引导人工智能发展始终围绕保障民生、提升就业质量发力。

构建适应智能时代的人才培养体系，加大人力资本投资，提高求职者技能和素质。将人工智能通识教育、数字素养纳入国民教育，注重人工智能原生人才培养。完善高等教育机构专业动态调整机制，及时更新专业课程设计，增设“人工智能+X”复合型专业，并将批判性思维、复杂沟通、领导力等能力培养纳入教育体系，与人工智能形成更强互补，高质量培养适应未来劳动力市场需求的专业人才。发挥企业、行业协会、院校等多元主体作用，建立终身职业技能培训体系，推行“人工智能+”模块化培训、微证书认证，满足不同层次劳动者的学习需求，让劳动者快速适应就业市场变化，不断增强就业竞争力。建设全国互认的人工智能培训体系，打通就业流动堵点。开展全民人工智能技能与素养提升行动，让更多普通劳动者掌握与人工智能共存的“元技能”。

构建多层次、高安全的社会保障体系，维护人工智能时代的民生底线。健全劳动法律法规与新职业规范，加强新职业劳动群体的权益保障。设计更加灵活的劳动规制规章，开发涵盖劳动关系、社会保障、能力发展、社会福利等领域的包容性政策工具，包括明确灵活就业、人机协作等新型劳动关系的认定标准；探索建立多层次社保机制，提升灵活就业人员的保障水平；优化参保缴费机制，利用大数据等手段拓展补缴、预缴等多种缴费方式，研究实行分险种参保机制，允许灵活就业人员选择参加单项险种社会保险。同时，健全职业伤害保障试点，开发适应新就业群体需求的职业伤害与医疗保险产品，全面增强劳动者的安全感。此外，健全人工智能规范与监管框架，明确其合法性边界，强化对新兴行业劳动关系的监管，避免就业算法歧视、非法信息采集等风险。

数字政务