CCID 赛迪智库 专报

2017 年 9 月 20 日　　第 59 期　　总第 555 期

人工智能监管：难点与渐进创新

【内容提要】　新一代人工智能产品具备决策自主性、自我学习和自适应能力，传统监管模式已无法很好地适应其发展需求，主要体现在难于预判智能产品应用的后果与风险，难以在问题发生后确定责任归属，并且难于在事前对潜在安全风险进行管控。赛迪智库规划研究所在总结多国人工智能监管经验的基础上，提出了我国加强人工智能监管的五点建议：对人工智能持包容态度，采取渐进式监管创新；建立安全标准与规范，明确安全责任体系；根据学习与适应能力，实现监管边界动态化；限制其自主度和智能水平，提升人们对智能产品的信任度；加入自我终结机制，防范系统性失控风险。

【关键词】　人工智能　监管难点　渐进创新

近年来，人工智能发展迅猛，国外有 Google、Intel、微软、苹果、特斯拉，我国有百度、腾讯、阿里巴巴等，这些企业纷纷通过自主研发、收购兼并等途径，跨界进入人工智能领域。新一代人工智能具有高度的自主性、自学习及适应能力等特征，给政府监管带来了新的挑战，比如，无人驾驶汽车的事故判定与归责等。显然，传统监管模式已无法很好地适应新的发展需求，甚至可能会对人工智能新技术和新模式起到阻碍作用。因此，需要适时根据人工智能技术发展阶段创新监管方式。

一、人工智能的监管难点

（一）难于对智能产品应用后果与风险进行预判

由于受限于自身的认知能力，人们在决策过程中很难分析所有与决策相关的信息，而通常是借助过去的经验，形成对未来相似事件的预先判断，进而选择自己较为满意的解决方案。然而，随着计算能力的不断攀升，人工智能可以计算出大量的可能性，能够轻易地去尝试那些人类以前从未考虑的解决方案，其选择空间往往大于人类。换言之，尽管人们设计了某人工智能产品，但受限于自身的认知能力，研发者无法预见其所研发的智能产品做出的决策以及产

生的效果。例如，AlphaGo 实质上是一个深度学习的神经网络，学习了近 16 万局共 3000 多万步人类职业棋手的走法，并且通过数百万次的自我学习，从中积累制胜经验，创造出在人类棋手定式之外的不同对弈策略，甚至连其研发者也难以预判 AlphaGo 将要采用的对弈策略。然而，基于深度学习技术的新一代人工智能产品大都具备自学习能力和自适应能力，如果放任自流，难以保证其行为结果与大众期望利益始终一致。Google 公司曾研发出一款智能数码相册软件，该智能产品在经过有监督的学习后，会把那些黑色皮肤人群识别为大猩猩，这就有可能引发种族歧视的争议。

（二）难于在问题发生后确定责任归属

人工智能产品一旦出现安全问题，划分责任归属可能会异常复杂，主要有以下原因：一是由于人工智能系统通常借助虚拟社区（如 Github 开源平台）进行协同研发，开发过程中可能会用到许多其他人开发的组件，数量众多的潜在责任人给权责划分带来了困难。二是大多数人工智能产品的内部运作并不透明，多数企业尚未公开其智能产品的源代码和测试信息，从而增加了监管部门确定责任归属的难度。三是许多人工智能产品在设计之初，便包含了诸多不受控

机制（比如后天自学习、自适应能力），一旦出现事故，一些法律灰色地带就给企业推诿责任带来了便利。2016 年 5 月，特斯拉 S 型电动轿车在开启 Autopilot 辅助驾驶模式下发生撞车事故，这起事故判责存在严重分歧，特斯拉公司、用户、传感器 Mobileye 公司各执一词，可能都需要承担一定的责任。2017 年 7 月，由 Knightscope 平台制造的一款打击犯罪的机器人在硅谷购物中心工作时出现自主决策失误，将一名男孩判定为危险人物，撞倒并开展攻击，导致该男孩受伤。事后，在责任判定与赔偿时，当事人家属、商场以及机器人设计公司发生了较大分歧。

（三）难于在事前对潜在安全风险进行管控

新一代人工智能的高度自主化特征可能造成人类难以预见的风险，加上人工智能研发过程较为分散和隐蔽，增加了事前监管的难度，如风险点监测和预警等。对于传统的潜在危险源，比如核威胁、环境污染等问题，由于这些潜在污染源需要建设基础设施、雇佣专业员工、购置生产设备等，其举动较容易被监管部门察觉，在确定危险责任主体时也不会出现责任不清晰状况。而人工智能的研发过程则与之完全不同。一是人工智能

研发所需的物理设备较少，一些研发人员可以租用弹性计算服务来训练人工智能系统，由于不具备物理可见性，监管部门发现危险源的困难程度大大增加。二是研发主体极为分散。由于开源技术的日积月累，普通大众借助个人电脑或智能手机就能完成具有特定功能的人工智能系统开发，这意味着潜在危险源极为分散。三是人工智能程序通常借助虚拟社区协同研发，系统由不同的组件构成，监管部门即便识别危险源后，也难以明确其责任主体。

二、国外人工智能监管经验

（一）重视人工智能安全监管，将其提升到战略高度

2016 年美国发布的《美国国家人工智能研发战略计划》中第四项战略即为“确保人工智能系统的安全”。它提出通过采取一系列措施，比如增强人工智能的可解释性和透明度，构建信任体系，增强可验证与可确认性，以保护人工智能系统免受攻击，从而实现长期的人工智能安全和优化。2016 年 10 月，英国科学和技术委员会发布了一份关于人工智能和机器人技术的报告，呼吁政府应介入对人工智能的监管，通过建立监管体系来

保障人工智能技术能够更好地融入社会经济，并产生符合人们预期的良好效果。英国政府试图在监管过程中引入人工智能技术，以增强监管的适用性，从检验和确认、决策系统的透明化、偏见最小化、隐私与知情权、归责制度与责任承担等方面，加强对人工智能安全性的管控。

（二）采取渐进式监管创新，确保监管规则的连续性

美国政府在无人驾驶领域采取渐进式监管创新，在许可颁布、无人车设计、驾驶系统等方面都制定了过渡性监管规则。例如，2014年10月，美国加州车辆管理局将29张自动驾驶汽车公共道路测试许可证，分别颁给了谷歌、戴姆勒、大众三家公司，获得许可的条件之一就是人要能够随时接管汽车。2015年12月，加州车辆管理局要求所有自动驾驶汽车的驾驶座上必须始终乘坐一名拥有驾照的人士，并要求汽车在设计方面必须具备方向盘、油门踏板、制动踏板等操控装置，以便车主在无人驾驶汽车系统操控失误时能够随时接管汽车。2016年3月，美国高速公路安全管理局（NHTSA）称谷歌无人驾驶汽车符合联邦法律，并且规定无人驾驶汽车司机可以是自动驾驶系统，而可以不是人类。

（三）增强行业自律，施加“人工道德”约束

目前，Google、微软等公司已在其内部设置了人工智能伦理委员会。太空探索技术公司（SpaceX）首席执行官埃隆·马斯克也于 2015 年底成立了人工智能非营利组织 OpenAI，试图通过开源开放预防人工智能可能带来的的灾难性影响，推动人工智能发挥积极作用。2016 年 6 月，Google 和 OpenAI 联合发布了五条人工智能定律，目的是为人工智能提供一个有效的行为约束，以使其不会在有意或无意中做出危害人类的事情。2017 年 2 月，马斯克、霍金等人连同数百名研究人员、科技领袖和科学家联名表示，完全支持人工智能应该在生产力、道德和安全领域遵守的23条基本原则，要确保人工智能为人类利益服务。

三、我国加强人工智能监管的对策建议

（一）对人工智能持包容态度，采取渐进式监管创新

人工智能在对经济社会造成巨大促进作用的同时，其存在的潜在风险不容小觑。政府对于人工智能领域的创新应持有包容态度，在迎接人工智能时代到来时，采取渐进方式进行监管创新，以确保监管规则的连续性。对于自动驾驶汽车，应采取

包容的监管态度，在试点示范中逐步探索规范相关领域产品、服务及安全标准，倡导企业自律和社会监督。英国政府就提议将汽车强制险的适用范围扩大到自动驾驶模式，利于驾驶者将汽车控制权完全交给自动驾驶系统时，其自身安全能得到保障。

（二）建立安全标准与规范，明确安全责任体系

为了保障人工智能产品与目标一致并能确保安全控制，监管部门对人工智能研发者应进行认证审批，强制要求其公布或提供与人工智能产品相关的安全信息，比如源代码、第三方测试结果等。在安全责任体系方面，政府可以从人工智能系统的开发者、生产者、销售者、使用者等角度，进行责任体系的设计。人工智能产品一旦出现安全问题，经过合法审批的开发者、生产者等将承担有限责任，而未经过审批的开发者将承担无限责任。用户在使用人工智能产品时，应遵守用户使用准则，如果由不当使用造成危害，用户也将承担相应的责任。

（三）根据学习与适应能力特征，实现监管边界动态化

由于人工智能产品通常都具备自主学习和适应能力，现有监管方法难以适用于不断进化的人工智能系统，监管部门应当

根据人工智能系统源代码，以及人工智能在测试环境的表现，从学习力、适应力等角度，对人工智能系统进行定期界定，判断其进化速度和所达到的程度，进而实现人工智能的监管边界动态化，使人工智能处于可控、安全的发展范围之内。

（四）在一定程度上限制其自主度和智能水平，提升人们对人工智能产品的信任度

为了提高人们对人工智能产品的信任度，限制其自主度和智能水平是有必要的，至少要让人们在心理上认为其拥有对人工智能产品的主导控制能力。一些传统装备（如汽车、高铁）虽然在速度上已经实现超越，但人们并不认为它是不安全的，因为人们对这些装备的认识和理解足够深入，认为能够驾驭得了，即拥有绝对的控制力。同时，建议根据不同领域特点，定向发挥人工智能的某项特定优势或技能。比如，让数据分析处理等能力成为人们辅助决策的工具，但最终决策权仍需掌握在人们手中。这也是当前阶段提升人们对人工智能产品信任度的关键。

（五）加入自我终结机制，防范系统性失控风险

人工智能最大的威胁是当前人类尚难以理解其决策行为，

未来失控的风险较大，而一旦失控则后果严重。正如衰老机制是内嵌于所有生命体中的必然，人工智能应该也存在自我毁灭机制。在这里，可借鉴 2016 年 Google 公司提出要给人工智能系统安装“切断开关”的想法，相当于在其内部强制加入某种自我终结机制，一旦常规监管手段失效，还能够触发其自我终结机制，从而使其始终处于人们监管范围之内，能够防范系统性失控风险。

本文作者：工业和信息化部赛迪研究院　　陆平　曹茜芮
联系方式：18811067149
电子邮件：luping @ccidthinktank.com

赛迪智库
面向政府 服务决策

研究，还是研究
才使我们见微知著

信息化研究中心	工业化研究中心	规划研究所
电子信息产业研究所	工业经济研究所	产业政策研究所
软件产业研究所	工业科技研究所	军民结合研究所
网络空间研究所	装备工业研究所	中小企业研究所
无线电管理研究所	消费品工业研究所	政策法规研究所
互联网研究所	原材料工业研究所	世界工业研究所
集成电路研究所	工业节能与环保研究所	安全产业研究所

编 辑 部：赛迪工业和信息化研究院
通讯地址：北京市海淀区万寿路27号院8号楼12层
邮政编码：100846
联 系 人：刘 颖 董 凯
联系电话：010-68200552 13701304215
010-68207922 13910685050
传 真：0086-10-68209616
网 址：www.ccidwise.com
电子邮件：liuying@ccidthinktank.com

报：部领导

送：部机关各司局，各地方工业和信息化主管部门及相关部门

编 辑 部：工业和信息化部赛迪研究院
通讯地址：北京市海淀区万寿路27号院南门8号楼12层
邮政编码：100846
联 系 人：刘颖　董凯
联系电话：010-68200552　13701304215
010-68207922　13910685050
传　　真：010-68200534
网　　址：www.ccidwise.com
电子邮件：liuying@ccidthinktank.com