►►►求解时代之问③

当技术挑战伦理，治理如何跟上？

本报记者　**王美华**

AI带来哪些伦理挑战

记者：当前，人工智能带来了哪些突出的伦理挑战？请结合具体案例说明。

中国国际问题研究院世界和平与安全研究所副所长张薇薇：人工智能引发的伦理挑战非常多元，比如隐私与数据安全、算法偏见与歧视、深度伪造与信息操纵等。但最值得警惕的，还是人工智能可能失控的风险。尽管各方都认同“人类必须主导人工智能”，但在哪些领域需要控制、控制到什么程度、以何种方式实现，国际社会迄今远未形成共识。

就国际和平与安全而言，最现实的威胁来自军事领域的自主武器系统——这类系统借助人工智能算法，可在无人直接干预或仅有限干预下，自主完成搜索、识别、跟踪、决策并攻击目标。目前，相关技术已趋成熟，并在情报收集、目标锁定等环节得到应用。目前AI自主选择目标并执行攻击的技术条件已经具备，且其效率远高于人类操作。各国虽声称攻击的“按钮”仍由人掌控，但问题在于，技术优势直接关联作战胜算，这极易诱使各方越过伦理红线。虽然将生杀大权交给机器，普遍认为绝不可接受，但军事行动高度保密，外界几乎无法监督，最终决定者是人还是机器，难以验证。因此，自主武器系统已成为当前国际和平与安全面临的严峻威胁之一。

中国信息通信研究院人工智能研究所所长魏凯：随着生成式人工智能能力持续增强，虚假内容泛滥、人类主体性弱化、智能代理引发情感依赖、生成内容无序乱象等多重伦理风险交织叠加，构成当下十分紧迫的治理难题。AI日益强大的内容生成能力催生严峻的信息真实性危机，能够低成本批量产出高逼真虚假文本与音视频，深度伪造门槛不断降低，被不法分子用于电信诈骗、操纵舆论，大幅提升真假辨识难度，持续扰乱公共认知、侵蚀社会信任根基。

与此同时，算法深度嵌入各类决策流程，使得人类主体性面临侵蚀风险，大众逐步习惯于将判断与选择交由机器承担，长期依赖容易造成独立思考与批判性思维退化，被动让渡决策权；情感陪伴类智能体广泛应用，依靠拟人化共情互动诱发过度情感依赖，模糊虚拟与现实边界，干扰正常人际交往，带来潜在认知偏差。

当AI犯错，谁来埋单

记者：当前，人工智能正从“动口”走向“动手”阶段，智能体开始自主决策、自主执行任务，这种转变给伦理治理带来了哪些前所未有的挑战？

清华大学人工智能国际治理研究院副院长梁正：当下人工智能发展的趋势之一就是智能体能够自主感知、分析推理、做出决策并执行任务，这给人的主体性带来了深刻影响。

其中，最突出的问题是权责分配。今年年初，“龙虾”智能体爆火，但随之出现了一系列问题。比如它自主执行了一些操作，有的甚至是误操作，删除了重要邮件，或者导致信息泄露。显然，这个责任不能由智能体来承担。使用者如果在权限设置上有疏忽，没有进行安全的授权管控，就可能导致这类问题。所以我认为，智能体监管或治理的核心，就在于责任如何明晰、权限如何控制。如果没有预先防范，在办公场景下风险或许还相对可控，但一旦进入自动驾驶、金融交易、医疗诊断等高敏感领域，风险就变得难以承受，这是当前最突出的挑战。

记者：当AI越来越多地参与自动驾驶、医疗诊断甚至公共决策时，人机之间的责任边界该怎么划分？

梁正：在智能体的使用和部署过程中，责任划分是最突出的问题。使用者固然要负主体责任，但提供者和开发者也必须承担相应的产品责任。办公智能体这类应用相对简单，但如果涉及具身智能、自动驾驶，责任划分就必须更加清晰。

首先，企业提供这类产品或服务时，必须有非常明确的告知义务。比如，此前一些自动驾驶厂商的宣传就存在过度渲染问题，明明是辅助驾驶，却给消费者造成可以放手的错觉。实际上，相关法律规范明确要求L2级及以下车辆驾驶者双手不能离开方向盘，但一些车企在宣传中有意无意地模糊了这一点。一旦发生事故，又说自己只是L2级别的辅助驾驶，怪用户没认真看说明书，这显然是在告知义务上留下了模糊地带。

对比来看，以L4级自动驾驶为例，系统以自动驾驶为主、人类接管为辅，权责划分就发生根本性变化：自动驾驶系统和车辆成为主要责任主体，一般情况下人不能干预车辆行驶。可见，不同类型和等级的智能体，在研发设计阶段就需按等级、分场景划分权责，并嵌入产品系统。将来，手术机器人也会面临类似情形，当它的能力提升到一定阶段，人可能只起辅助和保障作用，就像自动化产线，人不能轻易介入。因此，就智能体而言，责任绝不仅在使用者一方，设计者、开发者和提供者都须承担相应的产品责任。

记者：算法利用个人数据无形中塑造认知，使人误将引导视为独立判断。您如何看待这一现象？它对个体自主性、尊严及社会公平有何冲击？人又该如何彰显自主性？

梁正：这正是我们为什么强调教育的重要性。人工智能时代，学校不会消失，但教师的角色要发生深刻转变，要善于运用技术，教会学生如何辨别事实真伪。

人工智能对教育最根本的影响，是从传授知识转向教人做人。这不只是课堂上的说教，更需要在与同学相处、接触社会的实践中去体会。未来的学校教育一定是开放的，可能每周都要有一天走进社会，甚至直接去实习、去工作。

年轻人有年轻人的优势，一些不受规矩限制的创意，以前未必有机会实现，现在有了可能。所以，学校要培养批判性思维，帮人建立起对自身主体性的认知。我们现在的学校有思想道德课，但哲学和伦理教育还很不够，而这在智能时代非常关键。未来甚至审美素养、每个人的品位都可能成为一种重要资源。当这些素养丰富起来，人自然就有了批判性认知和鉴别力。接受过大学教育的人，基本判断力会强一些，但这种能力将来应当更早、更普遍地培养，这才是治本之策。

技术带来的影响，单靠短期管控很难彻底消除，最终要靠人自己有判断力、能主动质疑。算法推荐确实容易让人只看到想看的内容，为此我国出台了算法推荐管理规定，要求不能完全迎合用户偏好，要有随机推送，甚至故意推送相反观点。短期来看，这种应对方法可能有效；但长远看，单靠这种方式并不能让人真正走向理性。只有培养起判断力和批判性思维，人才能拥有主动性，即使面对相反意见也能主动鉴别，这才是真正的自主性。

治理怎样追上日新月异的技术

记者：目前AI治理面临哪些现实困境，又有哪些破解思路和方法？

梁正：治理困境首先要从风险来源入手。从机理上看，人工智能的风险主要来自3个方面。第一类是内生性风险，即技术系统自身特性导致的风险，比如失控。目前我们无法保证它完全按人的意图行事，今年以来，一些大模型已出现自主发现漏洞、攻击甚至自我迭代的苗头，将来还可能自我训练和进化，而我们目前尚缺乏有效应对手段。

第二类是应用层面的次生风险。例如，算法决策用于医疗诊断时，可能出现幻觉错误或数据偏差，给患者造成不可挽回的影响，这主要与技术的适用范围相关。而生成虚假内容属于技术的误用、滥用甚至恶用——同样的技术用于影视创作是好事，但用于造假、诈骗，那就变成了违法犯罪。使用不当甚至还可能导致极端风险，比如，对话大模型能力很强，但若缺乏防护，有人询问如何合成毒品或制造炸药，也可能造成严重后果。

第三类是衍生风险，与技术本身及具体用途关系不大，却带来广泛社会影响。当前最令人担忧的有两方面：一是大量智能体对人的学习、认知能力造成冲击，形成信息茧房、数字依赖，导致社交能力下降；二是就业问题，即人工智能对既有工作岗位的替代。这些本质上并非技术直接造成，而取决于我们如何处理与它的关系，如何调整社会结构去适应。

这三类风险的应对各有侧重。

针对内生风险，有两条破解路径。一是从基础研究入手，比如可解释人工智能、物理空间智能等，去解决技术的黑箱特征和可能失控的风险，让技术真正安全可控；二是从制度层面划定红线。正如基因编辑，在未知生殖基因编辑后果前，先划定禁区，等技术手段和认知成熟后再行推进。

次生风险主要靠制度解决。需要指出的是，现行法律框架已能应对许多问题，例如无论使用AI还是传统手段，凡侵犯人身权益或损害生命财产健康的，均属违法。对深度伪造等新问题，则可通过建立新的制度如内容标识、溯源管理等制度加以规范。

应对衍生风险，需要全社会更新治理理念，甚至重塑人机关系。智能革命已经催生人机协同、一人企业等新型业态，法律制度也需要做出相应调整，例如明确人工智能工作的法律属性、劳动被替代后如何保障人的工作空间等。

记者：您刚提到的可解释人工智能，是否就是用发展技术来解决技术带来的问题？

梁正：是的，必须让它变得可解释。现在大模型应用走在了理论前面，我们至今解释不了它为什么在理解、生成、推理上能做得这么好。虽已大规模使用，原因却说不清，只能归结为“涌现”。这方面的基础研究若能突破，我们就能说清这些能力的成因，包括幻觉的产生原因，进而有望从根本上解决问题。

当然，这个过程可能是滞后的。我常举飞机的例子：莱特兄弟的飞机早已飞上天，但人们在相当长时间内并不知道它为什么能飞。背后的原理如空气动力学是几十年甚至上百年后才出现的理论，此后才能据此设计真正安全的飞机。历史上，技术发展总是存在这种滞后，而解决内生风险的根本途径，或许就在于此。

记者：当前，技术发展日新月异，治理常常跟不上，怎么解决这个难题？2026世界人工智能大会主席声明提到“坚持风险导向，敏捷治理，探索分类分级管理”，如何理解这句话？

梁正：敏捷治理这个概念提出已久，主要是为了应对所谓风险社会，因为现代社会风险来源日趋多样和分散。过去环境相对静态，一套规则可长期适用，但现在已行不通。敏捷治理强调因时而变，并根据形势及时调整。同时，不能只靠单一治理中心，政府之外，社会、产业、公众都要参与进来，才能应对多样、快速变化且分散的风险。

对人工智能而言，强调敏捷治理，更是因为技术变动太快，治理往往滞后。因此，在人工智能治理上，我们一直强调“小步快走”。新问题出现时，不要一上来就管死，而应先研判现有治理工具是否够用。如果够用，可多考虑如何落实，并出台指引。例如，面对人工智能领域的争议，司法系统可快速出台司法解释，第一时间应对诉讼新问题。好处是不必等立法完备，因为到那时技术可能早已改变。

近几年，网信办在算法管理、内容生成、标识、安全审查等方面出台了数十个行政规范，快速回应共性问题并形成共识。待情况熟悉和各方认知趋同后，再出台条例，如《人脸识别技术应用安全管理办法》便是充分讨论后推出的。这就是小步快走，通过指引和规范持续迭代的思路。

目前，《人工智能安全治理框架》已迭代到2.0版，今年可能研制3.0，伦理安全指引也在更新。未来的思路是，针对不同应用场景不断深入细化，做到既不管死又不失规矩，既保安全又促发展。

如何促进AI朝向上向善的方向发展

记者：2026世界人工智能大会期间，世界人工智能合作组织应运而生，29个国家成为创始成员国；大会通过《主席声明》，国家发展改革委发布《人工智能合作发展行动计划》，工业和信息化部发布《国际人工智能伦理治理行动计划》，它们对全球AI治理的实际推动力将有多大？

梁正：人工智能已成为最大的全球公共产品，全球南方国家长期呼吁更公平的AI治理。在此背景下，中国勇于担当。一方面，中国的AI发展已走在全球前列，应用的丰富性和创新性得到广泛认可；另一方面，我们的实践正解决产业痛点和社会需求，例如中国气象智能预警方案“妈祖”已在巴基斯坦等7个国家落地，并在全球40余国“云端”应用。

过去3年，中国在全球AI治理方面主要推进了两项工作：一是在联合国层面推动能力建设普惠计划，通过培训班、共建创新应用中心等为全球南方赋能，这得益于我们在开源大模型及高性价比服务上的优势，始终以发展和赋能为本；二是始终坚持共商共建共享原则，强调开放共赢驱动创新、强化风险确保安全、包容并蓄文明互鉴、和衷共济完善全球治理。

当前全球治理碎片化严重，部分国家另起炉灶甚至退出多边合作，导致治理缺位。中国在此刻承担责任，明确强调在联合国框架下共同应对风险。新成立的世界人工智能合作组织，是全球首个人工智能政府间国际组织，按国际组织规则运作，是各国贡献经验资源的公共平台。中国可能是最大贡献者之一，但不改变平台的平等共建性质。至于推动力有多大，我认为关键在于后续运作。目前其共治定位已十分明确，为各国对话合作开辟了新渠道。若能持续细化规则、落实行动方案，将有效弥合治理赤字，推动全球AI治理从碎片化走向协同。

张薇薇：从倡议到组织正式成立，中国推动AI全球治理的思路越来越清晰。我们自身在产业实践和治理方面积累了丰富经验，同时坚定支持联合国发挥主渠道作用。目前，除美国明确反对多边治理外，大多数国家都持积极态度。在此背景下，中国发起组织、发布成果，就是希望在理念、原则上加强引导，分享实践经验，推动尽早形成具有广泛共识的全球治理框架。

有了平台，就能开展更多务实工作。当然，挑战仍然存在，其中最大的难题是如何将各国有效聚拢，开展真正有实质意义的讨论。世界人工智能合作组织现有29个创始成员国，可以先在这个范围内深入探讨。各国既有共通的基本原则，也有因国情和文化差异带来的独特性，需要通过持续对话弥合分歧，把共识落到实处。前路还有很多工作要做，但迈出第一步本身，就具有重要价值。

记者：尽管存在分歧，各国在应对AI带来的伦理挑战上仍有共同利益。哪些领域最有可能先形成共识与合作？

张薇薇：最容易形成共识的，是防范AI失控、避免人类被奴役这类极端风险，这是任何一方都不愿看到的局面。在恶意滥用方面，共识也相对容易形成，例如利用AI制造虚假或有害信息，恐怖分子或犯罪分子获取自主武器能力，甚至用AI开发生物武器等。至于系统性歧视问题，不同文化背景下的差异较大，这一问题在中国社会相对不显著，但在欧美较为突出，因此要形成共识可能需要更多的对话和协商。

记者：在人机协同、跨界融合、共创分享的智能社会愿景下，怎么才能确保技术始终服务于人的全面发展？如何促进全球人工智能朝着向上向善、造福人类的方向发展？

张薇薇：这涉及对“向善”的定义。全人类有共同的价值追求，不同文化也各有其“善”的内涵。我个人认为，一条可操作的衡量标准是：技术是否让人活得更轻松、更有价值感和尊严感，是否让人的潜能得到更充分的释放。举个例子，如果AI能帮我们减轻日常衣食住行的负担，人就能腾出更多精力投入到创造性活动中，去拓展生命的厚度和价值。与此同时，也必须守住底线——绝不能用技术作为伤害他人的工具。一旦借助工具对他人造成伤害，无论出于什么目的，都难以称得上是“向善”。

梁正：这关乎我们人工智能治理观的核心理念，也是与西方的重要分野。2019年，中国发布了《新一代人工智能治理原则——发展负责任的人工智能》，强调人工智能发展应以增进人类共同福祉为目标。只要坚持这一原则，发展就不会偏航。我们不认同西方某些企业先提升能力、出了问题再补救的做法，而是从一开始就强调：人人都应从人工智能发展中获益，人人都享有劳动权利，发展的终极目标是人的自由解放。技术把人从重复性事务中解放出来，是为了让人从事更有意义的工作，最终实现每个人的全面发展。因此，在技术开发阶段就要思考人如何发挥作用而非被替代；在使用阶段，则要更多审视它对人的意义和创造的社会价值。

记者：面对快速发展的AI，普通人应该如何理性看待和使用它？您有哪些建议？

张薇薇：首先要主动了解AI。很多时候，风险恰恰源于“不了解”。比如，西方曾出现青少年与AI对话后被教唆自杀的案例，国内也发生过AI换脸诈骗事件。但现在各种反诈宣传都在提醒大家：技术可以做到多么逼真。所以，普通人应当通过正规、权威的渠道，去了解AI能做什么、可能带来哪些危害，提前心里有数，就能大大降低风险。在此基础上，我们还应主动思考如何借助AI发挥自身特长，去做以前力所不能及的事情。当然，有一点必须时刻牢记——无论怎么用，都不能逾越法律的底线。

梁正：第一，积极拥抱，尽早使用。AI普及是大势所趋，排斥或回避只会错失机遇。就像互联网，谁先学习利用，谁就受益。我们一直鼓励学生，有工具就用起来，只有在实际应用中，才能切身体会它的优势与问题。

第二，审慎使用，始终以人为主。AI是辅助工具，人才是决策主体，主次不能颠倒。以科研为例，AI辅助研究日益增多，也带来造假等新问题，但我们不应因噎废食，而应主动将低创造性工作交由AI处理，如文献整理、调研纪要、PPT制作等。当AI能力进一步提升后，我们还可以与之互动，借助它拓展思路、提供备选方案，但它永远只是“参谋”，不是“替代者”。在此过程中，每个人都要持续反思：自己真正的优势在哪里？把长板做长，把不擅长的事尽早外包给AI。这种主动学习和适应的意识，是每个人都需要培养的。

如果AI代替我们来思考

王美华

“不少作业和论文的AI痕迹过于明显，这样下去，学生的独立思考能力可能会被削弱。”这样的担忧，正成为越来越多一线教育工作者的共识。如果思考“外包”给AI，长此以往，我们会失去什么？

技术为人类减负，本是文明进步的动力。从结绳记事到算盘，从计算器到搜索引擎，每一次工具的革新，都曾将人类从繁复的脑力劳作中解放出来。但生成式人工智能的颠覆性在于，它第一次直接介入并试图替代“创作”与“思辨”这种人类核心智慧。在拥抱这些工具时，我们或许需要停下来反思：我们外包出去的究竟是什么？

人的成长，精神的强健，离不开一个“啃”字。啃一本难读的书，在反复琢磨中豁然开朗；解一个棘手的问题，在山重水复中寻得柳暗花明。这个过程，看似缓慢，甚至痛苦，却是思想的根须往深处扎的必经之路。古人说“读书百遍，其义自见”，说的正是这个沉浸、咀嚼、内化的功夫。倘若跳过它，直取结论，就像吃别人嚼过的馍，能得到营养，却长不出自己的筋骨。知识的获取若太过轻易，理解就容易流于浅表，当求索的过程一再外包，我们可能会错过深入思考的契机，但思考本身如同肌肉，需要在反复的使用、撕裂与重构中才能强健。久而久之，世间怕要多出一批“知道分子”，少了许多真正的思想者。

AI擅长在已知中归纳、拼贴，却无法“无中生有”。它像一个向导，能带你走最平坦的路，但真正的创造，往往需要独自跋涉荒野，在迷途中遇见风景。回望人类文明的长河，那些照亮时代的灵光一现，往往诞生于看似“无用”的深思冥想。阿基米德在浴缸里悟出浮力定律，门捷列夫在梦中拼出元素周期表……这些闪耀时刻，是长时间专注思考的结晶，是潜意识对信息发酵重组后迸发的火花。倘若因为有了AI，我们就放弃深度思考，那些“顿悟时刻”，或许就会与我们擦肩而过。

笛卡尔说“我思故我在”，思考，是人确立自身主体性的根本方式。人类历史上每一次伟大的思想解放，都始于对既有观念的质疑与反思。然而，当人工智能以“客观”“全面”的面貌呈现信息时，完全可能用更隐蔽的手法编织偏见，让某种价值倾斜变得理所应当。在此过程中，独立思考的防线可能在潜移默化中被侵蚀。一旦失去这份警觉，对AI生成的结论全盘接受，对算法言听计从，我们交出的便不只有劳动，还有判断力、反思力与创造欲。

当然，指出这些，并不是要我们拒绝AI，而是要清醒地善用它。问题的关键，不只是技术，更在于使用技术的人以及背后的评价导向。荀子说：“君子生非异也，善假于物也。”AI理应是我们认知延伸的“杠杆”，而非代替自我的“大脑”。我们拥抱技术，是为了更好地去感受、去判断、去创造，而不是将这些人之为人的根本轻易托付出去。在时代的喧嚣中，唯有守住这份清醒、审慎与创造的定力，才能真正做到“善假于物”，而非“役于物”。