



技术进步带来虚假错误内容的生成风险

学习时报：近年来，生成式人工智能引起社会各界大范围关注。与以往内容生产主要依赖于人类自身的经验和创新相比，人工智能时代内容生成的核心变革体现在哪些方面？

薛澜：与以往人类社会的内容生产机制相比，从大数据中自动获取知识的机器学习是人工智能时代的主要机制和技术驱动力。随着大数据技术在多领域的广泛应用和市场规模的快速增长，生成式人工智能技术已成为人类进行内容生产的有力工具，拓展了新创意的展示方法，大大超越了人类以往的内容生产效率，甚至正在重塑人类社会内容生产和传播的底层逻辑。

人工智能技术对内容生产的颠覆性突破，首先体现在生产效率的指数级提升。比如，在传统内容创作中，一部作品往往需要很长一段时间，但通过大语言模型，只需要输入提示词或者初步构想，就可以实现秒级生成。其次，生成式人工智能技术的广泛应用在很多领域大幅降低了知识获取和内容创作的门槛，这不仅降低了专业创作的成本，更赋予了普通用户参与内容生产的能力，在一定程度上创造了“全民创作”的可能。

学习时报：对于各种生成式人工智能技术的开发应用，人们在享受便利的同时，也正在经受着各类虚假错误内容的困扰，您怎么看这类问题的产生？

薛澜：技术红利与技术缺陷是共生共存的。从生成机制上说，生成式人工智能带来的虚假错误内容主要分为三类。

一是人工智能技术本身具有的内生风险。比如，技术幻觉风险在内容生产过程中普遍存在。技术本身无法对源头数据进行真假辨别，加上理解能力有限、模型架构缺陷等原因，人工智能技术在实际应用中呈现出一定的内容重构特征，可能输出看似严谨合理，实际上却与事实或逻辑不符的虚假错误内容，出现“一本正经地说假话”现象。

二是技术开发过程中的风险。技术的算法歧视是由训练数据的思维映射而引发的。人工智能技术在训练、学习过程中所用数据，都是人类社会以往产生的，必然带有主观色彩和人类活动特征。生成式人工智能应用的过程，实质上是对过往人类社会模式的特征进行总结并用对未来社会进行感知和决策的过程，复制并延续传统社会特征自然也就不可避免，使得现阶段生成式人工智能技术无法完全避免歧视。整体而言，人类借助以往的经验、知识储备和信息查证，相对容易识别技术内生风险、开发训练风险引发的虚假错误内容。

三是基于技术应用或滥用的风险。任何技术本质上都是工具，它的伦理困境在于使用的方向问题。对这类风险要辩证看待。第一种情况，在善意的应用中带来了无意的负面影响。例如，网友用人工智能“复活”已故明星，本意是表达怀念，但未经家属同意，反而可能让逝者亲属再次承受痛苦，甚至涉及肖像权争议。技术应用初衷是善意的，但因缺乏伦理考量，带来了情感伤害和法律风险。第二种情况，深度造假与恶意应用。由于技术加持，主观恶意的内容造假突破了以往“耳听为虚、眼见为实”的真实性判断方式，语音合成、视频换脸等生成的“超真实”虚假内容冲击着社会信用基础，“眼见未必为实”具有以往谣言难以企及的破坏力，传统社会信用评判方式面临着崩塌风险。而且，恶意应用的危害力因技术的普及而成倍增长。更严峻的情形是，一些主观恶意的虚假错误内容生成已经形成完整的产业链条，从定制虚假内容生产到隐秘传播渠道，呈现出专业化、规模化的特征。与技术缺陷导致的虚假错误内容相比，这类内

人工智能时代内容生成的伦理治理

——访清华大学文科资深教授、人工智能国际治理研究院院长薛澜

□ 本报记者 王翠娟

容更不易被发现，危害程度更大。

所以，我们对人工智能技术进行监管，主要是建立一个合理的治理体系，对滥用、恶用人工智能技术所带来的各类已经产生的危害或尚未有实质危害的潜在风险进行规制。

人工智能时代需要敏捷治理

学习时报：由于生成式人工智能技术的飞速发展，传统治理模式已经跟不上现实内容生态治理需求，我们需要做怎样的调整？

薛澜：自2022年底ChatGPT发布以来，生成式人工智能技术日新月异，几秒钟就可以生成用户所需的视频、音频或文本，人工智能技术的超预期迭代速度和不可预测的发展特征，对治理体系的同步演进提出了巨大挑战。

一方面，以生成式人工智能为代表的人工智能产业目前属于新兴产业，其以创新为主要驱动力，满足并激发了海量的社会需求，可以说是一种在市场需求主导下通过技术创新和商业模式创新而形成的产业。高度不确定性是新兴产业的突出特征，关键技术的迭代，市场需求、竞争环境、商业模式、价值理念等方面的变化往往难以捉摸甚至瞬息万变，这就产生了技术发展与技术治理步调不一致的问题。同时，市场信心在新兴产业发展中起着关键性作用。一旦有关方面释放强监管信号，或发生有关产业发展的负外部事件，新兴产业的创新路径和方向将会受到直接影响，甚至遭受毁灭性打击。

另一方面，以生成式人工智能为代表的人工智能产业发展带来的伦理风险和负外部性问题也需要引起关注，并采取科学合理的措施降低社会风险，确保其健康可持续发展。在新兴产业发展过程中，政府监管能够对相关产业的技术和商业模式起到一定程度的正向引导作用，帮助企业有效规避可能产生的系列社会风险，营造负责任的企业形象，为产业高质量发展创造良好的社会支持氛围。

在人工智能时代，我们要坚持敏捷治理理念，做到监管不是关门禁入，而是实现开门有序进入。敏捷治理的目标框架主要包含两个维度，一是维护产业创新，二是规制产业发展带来的负面影响。具体地说，就是适应人工智能技术发展的速度要求，配置产业发展所需的制度资源，维护技术创新导向，防控产业风险，促进公共利益最大化，从而实现治理目标的有机平衡。

学习时报：这么说来，敏捷治理是具有一整套原则体系的治理模式。具体该如何实施敏捷治理？

薛澜：一是治理上的快速反应，发挥信号警示作用。传统治理模式的突出特征是治理对象的不确定性低，发展路径和方向比较稳定，政府有充足的时间充实论证、制定相关监管判断的依据，形成完整的制度化的规则体系。而以生成式人工智能为代表的人工智能技术瞬息万变，如果延续这种治理模式，可能尚未完全弄清监管判断标准，治理对象就已经发生了变化。以往我们因为理念调整不及时在其他领域是有过深刻教训的，必须引以为戒。所以，适应人工智能时代的产业发展需要，治理需要突出敏捷性，迅速对已经出现或有苗头性的风险进行规制，给各有关主体一个信号，并随着实践变化不断迭代，及时完善优化措施。

二是工具选择上突出柔性原则，引导为主，惩罚为辅。在传统治理中，政府监管是在发现问题、确定问题之后通过正常行政或法律程序制定实施规则的，政策工具以惩处为主，表现为“过程慢、力度大”。而对以生成式人工智能为代表的人工智能产业，敏捷治理在工具的选择上就要反过来，即“下手快、力度轻”。“快”可以减少企业技术路径和商业模式的调整损失；“轻”主要突出方向性和试探性，企业可以马上得知监管态度，知晓对产业治理的方向。需要注意的一点是，这种工具选择的逻辑是引导而不是惩罚，但如果企业对政府的有关引导视而不见、置之不理造成不良后果，政府也有必要采取严格的监管措施。

三是治理关系上实现重塑，形成共同体机制。在传统治理模式中，政府监管具有鲜明的问题导向，监管者与被监管者之间呈现典型的“猫和老鼠”关系。政府主要通过抓企业的负面行为实现监管，尽可

能制定一套全面完整的监管规则，而企业由于法规意识不强，为了扩大自己的盈利空间，会通过寻找规则漏洞来逃避监管，二者往往位于对立面。但在以生成式人工智能为代表的人工智能时代，企业很多时候不知道再往前会出现什么样的创新，政府也不知道会产生什么样的风险，监管者与被监管者对未来发展趋势处于“共同无知”的状态，二者之间的对立关系演化为互相依赖关系。这就需要双方在共同利益的基础上开展深入交流与充分沟通，探索找到一个合适的空间，使得企业的创新探索不会突破政府规制的底线，政府的规则不会完全地束缚企业创新的手脚。这样的合作关系有助于政府更好了解技术和市场发展走向，正确评估相关风险并执行合理的规制措施，也使企业清晰洞察产业发展可能带来的风险和政府治理目标，明晰产业发展边界，共同塑造和引领产业的技术路径和商业模式选择，实现政府和企业的合作共赢，满足良好的互动预期。

学习时报：在内容生成领域，实施敏捷治理的方向和原则已经很清晰了，但在微观层面，需要解决好哪些难题？

薛澜：敏捷治理要真正落地并不容易。在内容生成领域，我们面临的其中一个难题就是如何确定内容生成主体行为的责任以及采取什么样的惩罚措施。

理论上，实施内容造假的主体肯定存在违法甚至犯罪行为。但假如一个人生成了虚假错误内容，甚至严重一点说涉及敏感话题，但并没有进行广泛传播，也没有产生负面的社会影响，这种情况下如何判定行为的恶劣程度？假如造假内容不敏感，但造成了大范围传播，又该如何判定责任？如果用户仅是出于个人娱乐等目的生成了虚假内容，但被平台误推送，造成了不良社会影响，责任又该如何划分？这些都对政策工具反应的速度和灵活性造成了难题。

对虚假错误内容治理要形成多元合力

学习时报：由于生成式人工智能技术的低门槛、普及化特性，参与主体也比以往更为广泛和多元，政府、研发企业、应用平台、用户、普通大众等都是直接参与者和利益主体。刚刚讲了政府治理理念的调整，企业在内容治理中要承担哪些责任？

薛澜：在人工智能时代的内容生成生态中，企业作为技术开发和应用的主体承担着重要的责任。针对人工智能产业发展所带来的风险，我国已经形成了较为立体的治理体系，可以分为上、中、下层。上层是2019年发布的《新一代人工智能治理原则——发展负责任的人工智能》和2021年发布的《新一代人工智能伦理规范》。中层是针对人工智能具体领域的法律法规，比如2023年发布的《生成式人工智能服务管理暂行办法》、2025年发布的《人工智能生成合成内容标识办法》等。下层就是鼓励企业加强内部机制建设，成立伦理委员会等。不论是技术开发企业还是技术应用平台企业，都必须遵守国家相关的法律法规要求，确保生成的内容健康、准确、真实，并严格按照内容标识办法对人工智能生成的内容进行标注，确保消费者的知情权。

除了合规性要求，企业还需要多做一步，投入更多时间和精力做好技术自纠，防范和应对自身技术系统被恶用的风险。比如通过预训练数据清洗，对某些敏感数据进行有效识别并剔除，提高源头数据质量；在内容生成过程中设置过滤环节，及时识别用户的主观滥用、恶用意图，拦截违规生成内容；对医疗等敏感领域的内容生成实施人工审核兜底；通过用户反馈学习，拒绝违背法律法规、社会公序良俗的内容生成要求，降低模型伦理风险；等等。

学习时报：在内容生成生态的治理中，除了要发挥有效市场和有为政府的作用，我看到您的团队还提出了“有机社会”的治理理念。

薛澜：打造良好的内容生成生态，不仅要发挥政府、企业的作用，更需要形成利益共同体意识，建立多元治理格局，重视公众在内容生成生态治理中的地位和作用。

首先，公众具有用户和受众双角色，能够直观感知内容生成质量。要提高他们的辨别能力和维护良好内容生态、参与社会伦理判断的责任意识，对不良生成内容及及时反馈，必要时对各类虚假错误内容或恶意应用行为进行举报，配合有关部

门加强监督。

其次，使用者推动技术创新。随着新一轮科技革命和产业变革的深入发展，传统以生产者为中心的创新模式逐渐被以使用者为中心的创新模式所替代。使用者创新是指产品的最终用户基于自身需求和使用经验，对现有产品或服务进行改进或重新设计，甚至开发出全新解决方案的创新行为，其核心驱动力在于使用者对实际应用场景的深刻理解和对新需求的切身感知。生成式人工智能技术的研发与应用本身是满足市场需求的结果，而用户在使用过程中会由于新需求的出现，产生对现有技术的优化需求，从而对生成式人工智能技术进行需求导向的创新重塑。因此，我们应该积极鼓励公众广泛参与，发挥使用者创新的集体智慧，促进内容生成技术的功能优化和使用场景的创造性拓展。

最后，公众作为使用者也应加强自律意识，善用生成式人工智能技术，明确新技术新工具是用来创造价值、增加社会福祉的。在内容传播上要有责任意识，对接触到的虚假错误内容要坚决说“不”，从个体做起，汇聚推动人工智能技术良性发展的善意力量。反过来，为了提高公众对人工智能技术的认识和理解、对虚假错误内容的辨别能力，政府和有关机构要着力推动科普教育力度，建立信息验证平台，利用讲座、展览等方式普及人工智能技术的基本原理和应用场景，帮助他们认识技术的局限性和各类潜在风险。借助各类公众号、视频号等新媒体平台发布科普文章和视频，提升公众的科学素养和防范意识，培养公众对接收到的各类信息的正确态度和阅读习惯，为构建良好的内容生成生态营造良好的社会环境。

现在，很多机构在积极开发相关的内容监测平台，这方面存在很大技术潜力和积极社会价值。但事实核查有着巨大挑战，目前的治理存在着“道高一尺、魔高一丈”的困境，需要建立政府主导、产学研协同、公众广泛参与的立体化治理体系，形成强大的治理合力。

总之，创新和治理的平衡是人工智能技术发展的永恒主题。

内容生成技术会导致创造力依赖和情感钝化吗

学习时报：智能化一直以来是人类的梦想，但我们都没预料到它会以如此迅猛的速度成为现实。对生成式人工智能技术而言，未来我们需要警惕哪些新兴风险？

薛澜：随着人工智能技术的进一步发展，人类社会正在进入全新的智能化时代，未来生成式人工智能技术的发展可能会超乎我们的想象。可以肯定的是，未来生成式人工智能技术应用会更加广泛，更精准地理解并满足用户需要，生成更具多样性、个性化和创造力的高质量内容，为人类提供更加丰富的选择。这当然也会带来一系列的伦理风险。其中最值得警惕的是技术引发的人类创造力依赖和情感钝化风险。

首先，人工智能在内容生成、决策支持和日常任务中的广泛应用，可能导致人类逐渐丧失独立思考和解决问题的能力。例如，当人工智能能够完全替代人类完成写作、设计甚至决策任务时，无形中人类就减少了深度思考和反复试错的过程，而这些恰恰是创造力激发和积累的关键环节。如此一来，长期依赖人工智能生成的内容，可能会导致人类的原创思维和表达能力逐渐退化，最终导致整体创造力的萎缩。

其次，人工智能技术虽然提升了信息获取效率，但其中可能掺杂了用户算法分析和系统自身的算法歧视，从而将人们困在“信息茧房”中，限制了接触多元观点的机会，进而削弱思维的广泛性、灵活性和创新性。当算法不断迎合用户的偏好，人类主动探索、批判性思考和跨领域联想的能力可能会逐渐减弱，而这些能力正是创造力的核心要素。

再次，在情感层面，人工智能的介入可能改变传统的人际互动方式。例如，过度依赖人工智能聊天机器人来代替真实的人与人、人与社会的人际交往链接，不可避免会导致共情能力下降和情感表达简化。目前，人工智能技术已能够模拟人类情感并提供即时满足，而且部分群体对自我情绪价值的要求十分强烈，从而可能逐渐失去处理复杂人际关系的耐心和意愿。此外，人工智能在艺术创作，如诗歌、音乐、绘画等方面的广泛应用，可能使人类情感表达趋于标准化、模式化，削弱个体独特的情感深度和主观体验。

需要明确，人工智能技术的工具属性，在任何时候都不能改变。任何技术的发展，本质上都是人类出于科学角度对稳定性和精准性的追求，以弥补人类能力的不确定性短板。因此，在二者的关系中，人是居于第一位的，人发明并使用技术工具，决不能出现工具驾驭人的情况。人与工具的最本质区别之一就在于人的无穷创造力和真实情感感知力。从当前情况看，我们这一代人对此原则的认知比较清晰，但下一代，尤其是从童年时期即接触智能设备的青年群体，可能更习惯人机交互方式，对这种风险就需要警惕，就要进行积极干预和引导。

解决的关键就在于人类如何主动管理和平衡技术的使用。如果能够将人工智能始终定位为辅助工具而非替代品，鼓励人机协作而非被动依赖，同时重视教育体系中创造力与情感培养的核心地位，技术就可能成为增强人类能力的催化剂。以往的经验表明，技术进步往往伴随新的挑战，但人类的适应能力和主观能动性始终是化解风险的关键。因此，未来要建立合理的人工智能使用规范，确保技术发展服务于人类的长远福祉，坚决避免削弱我们的本质能力的现象。

营造有利于年轻人创新创业的良好生态

学习时报：我们谈论风险，本质上是希望塑造科技向善的理念，促进生成式人工智能技术在健康有序的轨道上发展才是我们的最终目的。

薛澜：对。内容生成技术也好，具身智能也好，它们都属于科技创新。我们中国的老百姓对科技创新的态度总体来讲是非常支持、积极拥抱的，接受程度普遍较高。与其他国家相比，这是非常难得的，也是我们的一个突出优势。

除了公众支持，我们还要创造一个更加有利于包括生成式人工智能在内的技术发展的良好创新生态。在人工智能领域，年轻人表现得十分抢眼，他们有想法、有热情，需要为他们提供尝试的机会。在内容生成领域，我们国家的监管是非常严格的，主要是为了保障国家安全、维护网络空间清朗、保护社会大众的个人合法权益，同时也在积极探索适应技术发展的监管模式。未来政府规制要增强包容性，明晰大的监管框架，在这个框架下推动形成浓厚的鼓励创新的氛围。

学习时报：这种良好的创新生态具体包括哪些内容呢？

薛澜：首先，要有一定数量的实力雄厚的研究型大学。一方面，高校是培养人才的，人才是创新创业的主体，活跃的年轻人很大程度上就是大学毕业生甚至在校。而且，实力雄厚的大学也会形成“人才磁场”，吸引国内外人才集聚。另一方面，高校是重大科技突破的策源地，可以通过校企链接，持续不断地促进科技创新成果转化，为企业保持科技领先地位、产业升级提供创新驱动力。所以，科研实力强劲的高校是基础。

其次，要形成多元的资本投入。在创新生态里面，资本的来源包括政府补贴、金融贷款，也包括民营资本、外资等等。企业对资本的需求是全周期的，如种子期需要高校创新基金、政府科研经费支持，初创期需要政府引导基金、风险投资等，重点是创新培育；成长期需要私募股权投资，上市后需要通过公开募股、企业债券等获取资本，重点是实现资本的规模化赋能；等等。只有形成丰富的资本结构，才能适应不同类型、不同时期的企业发展需求。

再次，比较适宜的生活环境。这个“适宜”涵盖很多方面。比如房价要友好，能提供多层次的住房保障，包括公租房、人才公寓等，让年轻人“有得住”，对安居有盼头；创业要友好，有完善的创业配套机制和创业氛围，对创业失败进行政策兜底，让年轻人“敢创业”；交通要友好，无论上下班通勤还是出差都十分便利，让年轻人“少折腾”；生活要友好，除了赚钱，还要满足年轻人对休闲娱乐的高品质精神追求，让年轻人“活得爽”；等等。

最后，打造无事不扰、有求必应的营商环境。“无事不扰”强调政府的定力，“有求必应”强调政府的能力，需要政府既克制大包大揽的冲动，又保持周到服务的热忱；既尊重市场规律，又不推卸责任，做好对各类企业健康有序成长的扶持和服务工作。