■理论探索

AI大模型：觉醒、恐惧与治理

□刘永谋

◆AI恐惧的特殊性在于其背后的AI拟人论思维。也就是说，人类很容易把AI当作人来看待，比如将Robot这个词翻译为“机器人”，偏离原初“机器劳动工具”的意思

◆就目前AI发展的实际状况而言，我觉得治理AI不宜过急过严，将长期主义与敏捷治理结合起来，通过平衡二者形成某种敏捷长期治理战略

2023年以来，AI大模型发展迅猛，掀起一波全球热潮。关注AI发展的人，已不限于AI的研发者、推广者，或者评论者、管理者，更包括深感生活正被AI深刻影响的普通公众。可以说，AI大模型已经成为公共性议题。

其中，AI大模型潜藏的风险、伦理与治理问题，也引起整个社会的广泛关注。对于AI的飞速发展，不少人表现出焦虑乃至恐惧的情绪，比如焦虑AI会让自己失业，恐惧AI统治人类。在类似情绪支配下，治理AI要从严从快的呼声越来越响亮。如何正确看待大家的情绪，应该如何治理AI大模型呢？

一、AI焦虑与恐惧

我认为，AI大模型的兴起，标志着智能社会开始进入AI辅助生存社会。在AI辅助生存社会中，人们的生活、学习和工作均在AI的帮助下完成。比如，文案写作会先交给AI完成初稿，接着由人进行调整、润色和提高；旅行者出门之前让AI规划几套游玩攻略，然后综合选择制定最终的出游方案；学习者将会有个人专属的AI agent作为AI教师，以响应个性化的学习需求。总之，我们很快要与AI共生。

接下来，很多人相信，也许不需要100年时间，机器人便有能力取代人类绝大多数的体力劳动和脑力劳动，AI辅助生存社会演化为AI替代劳动社会。显然，这将是生产力高度发达、物质产品极大丰富的富裕社会，一些人设想的“AI社会主义”“AI共产主义”值得憧憬。

伟大理想的实现，总伴随着各种问题、曲折和代价，智能社会的发展亦是如此。AI大模型至少对以下3个领域带来极大挑战：(1)失业问题，即它可能导致文案策划人员、原画师、工业设计人员、程序员、媒体从业人员和翻译人员等一部分脑力劳动者失业。(2)教育问题，即它可能冲击既有的教育科研系统，比如学生违规用AI代替自己做作业。(3)信息安全问题，即自动生成海量AIGC，真伪难辨、立场可疑，权属不清、追责困难，甚至挑战主流价值观和意识形态。

以失业问题为例。无论如何，人机劳动竞争与协作的局面将逐渐出现。但是，这种挑战究竟有多激烈，真的如一些人想象的会一夜之间发生吗？

2024年6月，百度公司无人网约车“萝卜快跑”在武汉走红，引发社会公众对于无人驾驶取代出租车、网约车的焦虑，一时间沸沸扬扬。结果甚嚣尘上一两个月，的士司机、网约车司机并没有大面积失业，萝卜快跑也没有停运。这说明什么？说明冲击确实存在，但速度并没有摁切换键那么快，大家的焦虑和恐惧过头了。

所谓AI恐惧，是社会公众普遍存在的一种对AI的忧惧情绪、拒绝态度与攻击行为，属于技术恐惧的一种。技术恐惧由来已久。比如，蒸汽火车刚刚出现的时候，英国的农民认为它让庄稼减产，让奶牛不再产奶。早些时候，边疆地区修公路、铁路，还有些人反对，理由是这会惊扰山神。可以说，技术发展史伴随着一部技术恐惧史。

AI恐惧的特殊性在于其背后的AI拟人论思维。也就是说，人类很容易把AI当作人来看待，比如将Robot这个词翻译为“机器人”，偏离原初“机器劳动工具”的意思。拟人论是非常古老的想法，与人类本身的思维方式有关。古代人相信泛灵论，山川河流花鸟鱼虫都有神灵。认为机器人是人，是古老意识形态在AI时代的新形式。

AI拟人化在三个层面催生AI恐惧：(1)认知上的拟人化引发“AI类人”恐慌；(2)情感上的依赖性引发“AI非人”恐慌；(3)信仰上对AI的崇拜感引发“AI超人”恐惧。AI触发失业恐惧，属于“AI类人”恐慌情绪。

我的意思并不是萝卜快跑没有任何问题，而是大家恐慌过头了。面对这些实际的冲击，的确也需要我们进行整体的应对。面对社会舆论的质疑，专家、政府和车企应该加强与社会公众的沟通和交流，妥善处理已经暴露出来的问题，为新科技应用、推广和落地保驾护航。

二、AI觉醒与宣传

AI恐惧与AI企业、AI媒体有关系吗？有。AI拟人论流行，与AI宣传术有关。

所谓“AI宣传术”，即AI传播过程中面向社会和大众所采取的技术推广策略。影响AI公众印象的，主要是AI公司和与之关系密切的大众传媒。为了获得社会关注，AI宣传术常常对AI进行夸张推广，通过炒作吸引更多资金投入。

AI宣传术主要包括三种策略：(1)科幻叙事，即不仅科幻色彩浓厚，也注重用科幻文艺进行宣传；(2)偶像叙事，即马斯克、奥特曼等企业主亲自代言，扮演“真人超级科技英雄”，为AI“圈粉”；(3)觉醒叙事，即借助“AI奇点正降临”“AI已有意识”“AI可能统治人类”等“AI觉醒”类话题炒作，AI拟人论日益流行。

自从AI概念提出之后，在每一波AI热潮中，“AI觉醒”屡屡被炒作。1997年，Deep Blue击败俄罗斯国际象棋大师，有人觉得它有意识了，甚至还编造它输给人类后恼羞成怒放电杀死对手的假新闻。2016年，AlphaGo大败围棋世界冠军李世石，又有人认为AlphaGo有意识了。之后，ChatGPT也被某些工程师认为有了意识，DeepSeek则被很多人认为情商很高。

AI觉醒了会如何，超级AI到来会如何，超级AI有没有道德，有没有欲望，有没有目标，会不会统治人类……类似问题具有明显的幻想色彩，因而成为科幻作家趋之若鹜的切入点。AI科幻文艺有力的宣传，使得AI圈子具有很强的科幻气质和大众娱乐气质。

早在20世纪90年代，就有一些学者如罗斯扎克指出，AI觉醒、超级AI是AI圈子“吸金”“吸睛”的“法宝”：通过类似问题严重不实的娱乐化讨论，吸引资金流向AI从而壮大自身。

应该说，AI宣传术对AI发展的确起到很好的推动作用。原因起码有三：第一，它抓住人们的“痒点”。很多人会觉得类似问题很好玩，很有趣，特别适合白日做梦，适合成为娱乐元素。第二，它抓住人们的“痛点”。宣传超级AI马上到来，很可能统治人类，可以激起AI恐惧。最后，它抓住社会的“热点”。比如，气候变化问题讨论很热，有人宣称气候变化问题只需要有了超级AI就能解决，因为超级AI计算一切、无所不能，于是气候热点也被纳入AI宣传术当中。

当然，AI宣传术自然也导致一些问题。比如，不实宣传现在无人驾驶可以完全交给AI，可能导致事故；过度宣传AGI、超级AI，让人担心智能机器伤害我们。因此，要反思目前的AI传播方式，吸引社会关注又不挑起过度的AI恐惧。

三、AI的文明危崖

在各种AI恐惧中，最为深沉的是AI文明危崖问题。

何为文明危崖？有人指出，20世纪下半叶新科技发展，让单个人有能力实施危害巨大的恐怖行为，巨大灾难可能是由无心的技术错误引发。21世纪以降，很多思想家开始关切：当下的文明是否存在全局性的崩溃，甚至灭绝的生存性风险，使得人类社会如跌下悬崖一般，突然陷入黑暗甚至永夜之中？这便是近来全球广为讨论的“文明危崖问题”。

AI至少有两个问题直接与文明危崖相关：第一，AI特别是超级AI会不会总体上威胁人类文明？第二，对于全局性生存性风险的应对，AI是否能够有所助益，还是只能加剧潜在的危险？

AI可能导致的文明危崖风险包括两类：(1)AI灭绝，即AI可能灭绝人类；(2)AI衰退，即AI可能导致文明衰退。只有越过AI文明危崖，才有机会奔向“数字共产主义”。

AI灭绝与超级AI的出现直接相连。只要超级AI真的出现，AI灭绝的风险就很大。随着人工智能的不断迭代升级，可能会在某一时刻越过“奇点”，出现某种超级机器意识，不甘于做人类的劳动工具，而是要翻身做主人，甚至因为某种原因比如争夺资源，而灭绝整个人类。

超级AI如果真的出现，人类很可能无法揣度它的意图和目标。在《超级智能》中，尼克·波斯特洛姆设置了一个思想实验：假设有一家回形针工厂买了一台超级智能计算机，人类主管给它下达了一个看似简单的任务：生产尽可能多的回形针。这台超级智能计算机为了达成“生产最多回形针”的终极目标，开始不择手段。它意识到生产更多回形针需要电力、钢铁、土地等资源，而人类不太可能会放弃这些资源，于是为了扫除障碍，它选择征服整个地球，杀死所有人类。最终可能将整个宇宙的物质都变成回形针。

AI衰退往往指向“AI机器国”，不一定需要超级AI。所谓“AI机器国”，可以类比为一架严密的智能大机器，每个社会成员都成为其中一个小小的智能零件，随时可以更换，和钢铁制造的零件没有差别。AI机器国是一座大监狱，无处不在的监视、无处不在的控制充斥其中。从本质上说，AI机器国反人类、反人性，将AI用作操控人类的工具，导致文明陷入黑暗之中。

在《生命3.0》中，迈克斯·泰格马克设想，超级AI在灭绝人类的过程中可能会留下少数人，作为研究对象关在动物园中，与其他动物共同展览。在此种情形下，人类会逐渐退化成动物，文明不可遏制地不断衰退。

总的来说，虽然在流行文化尤其是科幻文艺中，AI文明危崖问题非常吸引眼球，但严肃的思想家并不太关注，因为不少人认为“文明的AI危崖”远比不上气候变化、核大战和新发未知病毒等的威胁。

但我认为，AI的生存性风险虽然很难排在前面，但AI与全球性灾难融合在一起，将极大地增加文明危崖的风险。比如，AI+核大战，AI+生化武器，AGI正在加剧能源危机等。再比如，AI与权力结合，用AI改造智人的身心设计不当，都可能导致AI衰退。

四、敏捷长期治理

文明危崖问题之所以会出现，从根本上源于对新科技失控的担忧。AI的发展可不可能失控？可能性当然存在。

人文研究应关注新科技发展的社会冲击，尤其要预测风险、预见问题，提醒社会未雨绸缪，防患于未然，看护社会福祉和公众利益。发展AI应该有所为、有所不为，对AI发展进行全面的、长期的选择、引导和控制。如果怀疑AI的某一发展方向非常危险，就应该停止、转变和重置此种AI发展进路，此即我所称的“AI发展的有限主义进路”的基本思想。

比如，如果不能确认超级AI的安全性，人类就应该果断选择避免AI产生意识的技术发展道路。比如，将规则制定、价值选择始终牢牢抓在人类手中，对AI实施价值对齐，因为只有人类能承担所制定的规则导致的责任和后果，而AI负责听命于人类，执行人类的指令即可。

“AI发展的有限主义进路”并非以人文为名阻碍新科技发展，而是以新科技健康发展为目标的保驾护航之举。新科技的发展短期冲击会被人夸大，而长期影响却被忽视，治理AI应该坚持长期主义战略。

这要求我们深入研究智能革命对当代社会的巨大冲击。AI有效加速主义者强调人机共生，但意识不到人机共生并不天然地等于人机共赢。这其中需要我们经过慎重的审思。

目前，AI大模型的发展突破不小，也不要吹得神乎其神。AI宣传术能要来钱，但陷入AI恐惧的公众可能要求国家和政府加以强力干预，结果是各种敏捷治理的想法出现，一些企业感到太严受不了。过度宣传有没有责任？肯定有。

AI相关政策的制定者和决策者，要对各种AI话语进行仔细的审度，分辨它们的言说者、言说场域、言说方式、言说动机以及主题流变等，准确而恰当地定位AI发展的真实状况。

就目前AI发展的实际状况而言，我觉得治理AI不宜过急过严，将长期主义与敏捷治理结合起来，通过平衡二者形成某种敏捷长期治理战略。

社会要加强AI科学传播，提高大家的AI素质。如此，社会公众面对AI宣传术就会有自己的判断，不会轻易被AI恐惧牵着鼻子走。

当然，不能说AI恐惧完全没必要，要彻底根除。辩证地看，AI恐惧让社会关注AI发展，尤其警惕AI发展中可能出现的负面效应，及时评估风险，并采取恰当的应对措施。但是，类似情绪要适度。

总之，AI的发展与每一个人都有关，如果人人都来关注AI，参与到AI向善的行动中去，智能社会一定能拥有更为美好的未来！

（作者系中国人民大学吴玉章讲席教授）

■综 述

城乡互融，诗歌何为

本报记者 黄尚恩

5月3日，一场以“城与乡的诗意重构”为主题的研讨会在重庆卫盾农场举行。卫盾农场位于重庆市璧山区健龙镇玉林村。一走进农场，“种田、写诗，都是大地的语言”“慢下来，等一朵云从山那边走过来”等红色条幅映入眼帘；几只蝴蝶在翻飞，有的闻闻架子上的葡萄花，有的嗅嗅青黄相间的枇杷果；倾耳听，似乎有人正在进行网络直播，向远方的朋友介绍来自大地的诗意……这样的情景，与“城与乡的诗意重构”的研讨主题，正好暗合。

从“乡土中国”到“城乡中国”，中国的“城”与“乡”始终处于不断融合的进程之中。在《作家》主编宗仁发看来，乡村城镇化和城市乡村化是现代化进程中一体两面的事情。过去的一些乡村，无论是房屋居所还是卫生条件，都不是那么宜居，所以要照着城市的标准来进行改造。城市建设也不能都是钢筋水泥，而是要像乡村那样，把房屋建在花草树木之间，建设花园城市、生态城市。城与乡的差异在慢慢缩小。

城与乡的诗意重构，一方面是现实层面的，即城市与乡村在建设中相互借鉴，彼此融合；另一方面是文学层面的，就是用诗意的方式去表达城乡互动中人的情感交流。上海交通大学教授何言宏表示，在这个过程中，需要尊重不同主体的独特性。比如说，我们要建设“文化农场”，不仅仅是要用城市文化来赋能农场，还要让农场生长出自己的文化，并使之对当下城市生活产生作用。

“书写乡村，特别是关于‘乡愁’和‘田园牧歌’的书写，如果被赋予过多虚拟的理想化的色彩，就有可能演变为一种与现实相脱离的文化想象。”诗人、《草堂》诗刊主编梁平谈到，进行城与乡的诗意重构，首先是要深刻认识城与乡的真实状况，包括它们之间的现实差异以及生活在其中的人的不同情感结构。现在城里的人想到乡村去，乡下的人想到城市来，本质上是人们对城乡心理落差的主动调适。因此，关于城乡的文学书写，绝不是简单地去描摹城市与乡村的表象，而是要聚焦人与人之间的有效互动。在这种互动中，人们有了对彼此处境的深入了解，进而真正实现情感的融合与共鸣。

“在对城乡进行诗意重构时，我们需要丢掉以往的对于乡村的‘歧视性叙事’。”诗人臧棣谈到，在过去一段时期的城乡叙事中，城市往往代表先进与文明，乡村往往代表落后与愚昧。后来，我们对这种观念进行了反思。实际上，在乡村世界中，人与自然大多处于一种和谐的状态。城市中的个体来到乡村，生命中那些压抑的记忆和感受无形中就会被释放掉，感到身心的放松。乡村世界当然也不是完美的，也存在一些问题。但是，从文化观念上，我们需要反思那种“以城市衡量一切”的态度。清华大学教授西渡同样谈到，乡村的一些生活哲学，恰恰是我们改造城市生活时需要借鉴的。比如，在乡村，万物相互依存，不断循环利用，构成了一种生生不息、永远在生长的状态。而我们的城市生活，还有诗歌写作，恰恰需要借鉴这样的精神与状态。

推动城与乡的现代化发展，既要把人们的生活环境搞好，也要推动人的思想观念的现代化。诗人阿尔丁夫·翼人谈到，无论生活在城市还是乡村，我们都需要解决人的心灵的问题。在这个过程中，文化发挥着重要的作用。有些人根本不写诗，但我们会觉得这个人很“诗意”；而有些人写了很多诗歌，但我们有可能会觉得他是个“俗人”。这就是要看，一个人，有没有自己的精神格局。诗歌评论家庄伟杰说，“诗意地栖息”，就是要让人回到本真的状态。当然，任何个体也不可能完全脱离社会环境而存在，这就需要以出世之心做入世之事。对于诗人而言，要持续锻造自己的心性，追求创作的“高品质”而非单纯追求“量的累积”。

书写城乡的互动与交融，需要诗人进行观念和技法的更新。北京师范大学教授张清华认为，新诗诞生以来，一直面临着如何将现代城市经验有效诗化的课题。当然，现在的乡村也不再是传统的乡村。乡村生活条件的改善，促进了乡村人精神世界的变迁；反过来，这种观念的现代化，也促进物质世界的发展。这些都需要我们不断提升以诗歌处理现实复杂状况的能力。中央民族大学教授敬文东表示，五四以来，诗人们用现代汉语来书写现实生活，特别是表达城市的复杂状况。到了当下，城市与乡村又面临深度的重构。如何让人们真正活得诗意，如何以诗歌表达出这份为诗意而进行的努力，都是值得我们不断探索的课题。

此次活动由北京师范大学当代文学创作与批评研究中心、山东大学诗学高等研究中心指导，重庆市璧山区作协主办，重庆卫盾农业文化发展有限公司承办。

■书 评

知人论世与批评的有效抵达

□郝敬波

如何进行有效的文学批评一直是学界面对的重要问题。所谓有效，简而言之就是能够实现对作家作品阐释的预期目的，而这个预期目的之一就是成功抵达作品的艺术世界。如果文学批评没有找到通向研究对象的准确通道，仅在外围徘徊，或找到通道但浅尝辄止，那么这种批评就很难说是有效的。徐则臣作为“70后”作家的代表，一直受到批评界的关注。李徽昭教授新著《到世界去：徐则臣小说及其时代》以有效的批评方式引领读者深入抵达了徐则臣宏阔的小说世界。

“知人论世”的批评方式是该书最为重要的特征。“知人论世”是孟子提出的批评原则：“颂其诗，读其书，不知其人，可乎？是以论其世也，是尚友也。”该批评原则长期影响了中国文学批评。“70后”作家是在中国社会加速发展的过程中成长起来的，他们的书写对象也深刻反映了时代的变化和转型。李徽昭敏锐地抓住了这一点，书名“到世界去”所蕴含的不仅是徐则臣书写的重要内容，也是中国社会发展的时代特征。“到世界去”是徐则臣小说的重要主题，对这个问题的讨论是走进其小说世界的关键入口。李徽昭在该书第一章就以“中国与世界”为题，对徐则臣小说世界的形成场域进行探讨。上世纪八九十年代，中国作家开始更多地走向域外，由此获得了前所未有的“世界”体验。随着域外交流的逐步深入，徐则臣等新一代作家的世界视野进一步扩大，他们的世界体验、世界意识也融入到生命体验中，并深刻影响了小说世界的建构，正如书中所写：“以徐则臣为代表的年轻一代作家，他们的世界观、文学意识、文学创作，特别契合90年代之后的时代变革与全球化状况，既试图以本土视角与世界文学发生多元共振，又毫不丧失中国主体性，因此其审美观念呈现出一种新的世界性视角，以审美共通性的小说文体不断与世界进行对话。”

更为重要的是，李徽昭从“知其人”的视角出发，对徐则臣“到世界去”的独特文学品格进行了深入分析，指出徐则臣的小说在世界视野、全球问题等方面体现了更为阔大深远的眼光，却又极具本土性。李徽昭分别从“文本的世界表意与思索”和“中国传奇如何接通世界”两个维度讨论“到世界去”书写的个性化特征，探讨徐则臣在创作中发育和生成的艺术经验。李徽昭认为，徐则臣的“到世界去”的书写区别于一般的认识论意义，而是带有“思想性”的对话特征，“用中国方式对西方焦点空间及其文化展开了全球想象与本土再造，借此对世界文学空间进行新开拓，强化了与世界的思想对话”。而且，在这个过程中，徐则臣表现出在世界视野中建构中国文学的努力，在世界文学的交融中实施“本土文化面向世界的主体穿越和重构”。可以看出，李徽昭在“知人论世”的批评原则下找到了通向徐则臣小说世界的重要入口，对“到世界去”这一审美特征进行了极为深入的阐释。

在“到世界去”的阐释基础上，李徽昭以“推源溯流”和“沿波讨源”的讨论方式，在多元视阈下对徐则臣的小说世界进行了深度的细微探析。“形象之辨”部分重点讨论了徐则臣小说中的女性形象和知识分子形象，并指出其独特的审美价值。“意象书写”部分讨论了食物、船等意象，重点探讨了意象的古典审美意蕴以及徐则臣对意象诗学的探索。“长篇、时代与情感”部分则进行了代表作品的细读，对“花街”“京漂”和“谜团”等系列的小说进行了细致分析。对《北上》的分析是该部分的重点，在系统分析文本的基础上，归纳该小说之于当下长篇小说创作的价值与意义，特别指出“《北上》与世界和历史进行了多重对话，在题材和形式上拓宽了长篇小说的写作版图”。接下来是“文本短论”，对《王城如海》《青云谷童话》《跑步穿过中关村》《午夜之门》等小说进行解读，拓展了对徐则臣小说世界的观察边界。在以上的研究中，李徽昭拒绝那种面面俱到、四平八稳的讨论方式，而是建立在自己阅读感受的基础上，由感悟出发，靠近文本，然后进行思辨性的探析，从而在多元视阈中实现了对小说世界的深刻阐释。

该书虽然不是按照一般“学院派”的论著来结构的，但是其主要部分依然是思辨性的分析。在没有破坏这个研究基调的基础上，李徽昭增加了关于徐则臣创作的相关访谈及交流的电邮内容，以及与论题内容相关的通讯报道、随笔等别一向度的文本，这赋予了该书很大的灵动性和丰富性。而且，这些介入的元素与思辨的分析形成了鲜明的互文效果，相互渗透，相得益彰，从而丰富、深化了读者对徐则臣小说世界的认知和理解。

（作者系江苏师范大学文学院教授）