“深度学习三巨头”之一杰弗里·辛顿请辞谷歌，警示人们人工智能风险紧迫

新技术总会伴随新风险，AI为何更棘手？

■本报记者 沈湫莎

站在“人类文明的十字路口”，AI何去何从？这是摆在人类面前的一道必答题。自去年底ChatGPT横空出世、5天突破百万用户以来，有关AI与人类未来的讨论愈演愈烈，从产业界到学界，至今尚无定论。

回望历史，科学技术的发展总是在反对和质疑声中一路前行。新技术总会伴随新风险，这一次为何更棘手？作为地球文明的主导者，我们似乎遇到了一个前所未有的挑战：以目前AI的发展速度，人类会否沦为硅基智慧演化的一个过渡阶段？这一次人人都是参与者，无人可以置身事外。

——编者

继全球千余名业界大佬公开签名呼吁暂停研发比GPT-4更强大的人工智能（AI）系统后，在业内有“AI教父”之称的计算机科学家杰弗里·辛顿（Geoffrey Hinton）本月初宣布离职谷歌，进一步加深了人们对AI是否已经失控的担忧。

现年75岁的辛顿在神经网络领域长期从事开创性工作，为AI技术的发展奠定了基础。在3月下旬那封聚集了1000多个签名的公开信发布时，“深度学习三巨头”、三位2018年图灵奖得主表现各不相同。其中，蒙特利尔大学教授约书亚·本吉奥（Yoshua Bengio）高调签名，脸书首席人工智能科学家、纽约大学教授杨立昆（Yann LeCun）旗帜鲜明地反对，身为谷歌副总裁的辛顿一言未发。

宣布离职谷歌后，辛顿公开表达了自己的观点。他在接受《纽约时报》采访时表示，相较于气候变化，AI可能对人类的威胁“更紧迫”。他甚至表示，之所以离开谷歌，就是为了完全自由地说出AI潜在的风险，向世人提出警示。在5月3日麻省理工技术评论举行的一场半公开分享会上，辛顿坦言，过去他认为AI风险是遥不可及的，“但现在我认为这是严重的，而且相当近，但停止发展AI的想法太天真了。”

人类又一次陷入“科林格里奇困境”

人类历史上，曾无数次因新技术的诞生而产生担忧。比如，1863年瑞典化学家诺贝尔发明硝化甘油炸药，1885年德国人卡尔·本茨和戴姆勒发明内燃机汽车，20世纪40年代人类发明了原子弹……

辛顿如今对AI的情绪，与晚年时期的诺贝尔和爱因斯坦如出一辙。诺贝尔发明炸药原本是为了提高开矿、修路等工程的效率，当他发现自己的发明被投入到战场上后，其自责达到了顶点，这也促使他后来创立了诺贝尔奖。

清华大学人工智能国际治理研究院副院长梁正认为，在新技术兴起时，人类通常会面临所谓的“科林格里奇困境”——技术尚未出现，监管者无法提前预测其影响并采取行动；而当它已经出现时，采取行动往往为时已晚或成本过高。

幸运的是，迄今为止，人类通过不断适应新技术的发展而调整治理手段，一次次走出了“科林格里奇困境”。“比如在汽车大规模上市前，人们就为其安装上了刹车，之后也一直在完善汽车的安全性，为其提供各类检测和认证，甚至在汽车发明100多年后，人们还在为其安全‘打补丁’——装上安全气囊。”梁正说。

不过，他也承认，这一次AI的来势汹汹似乎与以往有些不同，“速度太快了。”比如，训练了几个月的ChatGPT的性能提升比过去几年迭代都要快，这意味着生成式AI大模型可以在几周内将潜在风险转变为实存风险，进而对人类社会造成不可估量的影响。

另一个不同是，这一次我们似乎无从下手。辛顿坦言，对于气候变化的风险，人类可以提出一些有效的应对策略，比如减少碳排放，“你相信这样做，最终一切都会好起来的。但对于AI的风险，你根本不知道如何下手。”

另外，商业竞争也会促使大模型一路“狂飙”。OpenAI深知GPT所蕴含的潜在风险，尽管他们对外表示将审慎推进AI系统的研发，但并不愿意就此暂停或彻底放慢脚步，而是期望社会为此做好准备。今年2月，这家公司刚刚公布了其发展通用AI的雄心与策略。而其首席执行官山姆·阿尔特曼表示，通用AI在AI技术上只能算是一个小节点，他们的远景目标是创造出远超人类智能的超级AI。

AI真的拥有人类智能了吗？

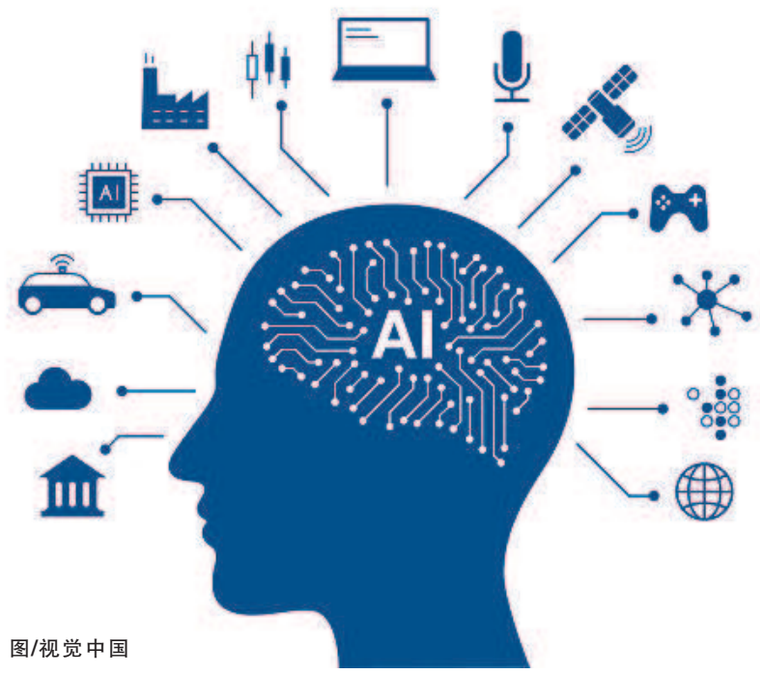
AI失控的故事，一直出现在科幻小说中。在大模型出现前，人们也对AI保持了相当警惕，但从未像今天一样如临大敌。那么，能识别出照片中的种种不合理、在各项考试中拿到高分、与人如沐春风般对话的大模型，真的已经拥有人类智能了吗？

联合国教科文组织AI伦理特设专家组专家、中国科学院自动化研究所AI伦理与治理中心主任曾毅认为，以ChatGPT为代表的大模型是“看似智能的信息处理”，与智能的本质没有关系。

“人们之所以觉得它很厉害，因为它的回答满足了人们的需求，如果这些回答来自于一个人，你会觉得他太聪明了。但如果你跟它说‘我很不高兴’，它说‘那我怎么能让你高兴一些’，这让人觉得它似乎理解了情感，而实际上它只是建立了文本之间的关联。”曾毅认为，目前的AI系统与人类智能的区别在于，大模型没有“我”的概念，没有自我就无法区分自我和他人，就无法换位思考，无法产生共情，也就无法真正理解情感。

清华瑞莱智慧科技有限公司AI治理研究院院长张伟强表示，当前的大模型仍属于第二代AI，其主要特征为以深度学习为技术、以数据驱动为模式。这使得它在决策链路和逻辑上具有天然的不确定性，即便是模型开发者，也无法准确预知模型的输出结果。当语言大模型“信口开河”时，不能认为是出现了所谓的“自我意识”，而仅仅是技术缺陷使然。

“计算能力当然是智能的一种，但智能的范围比这大得多，除了计算还有算计。”梁正说，如果把智能看作一个球体，阿尔法狗表现出的智力如同一个针尖般大小，大模型则是球体表面那一层，离真正的智慧内核还差得远。因此，许多技术派将ChatGPT视作某种“高科技鹦鹉”或人类知识库的映射。他们并不认为情况已经十分危急，因为与人类智能相比，大模型并没有触及认知的底层逻辑。



图/视觉中国

大模型真正的威胁在哪里？

我们应该允许机器充斥信息渠道传播谎言吗？应该将所有工作（包括那些让人有成就感的工作）都自动化吗？应该去开发可能最终超越甚至取代我们的非人类智慧吗？应该冒文明失控的风险吗？细读那封千人签名支持的公开信不难发现，业界大佬们并非为AI的智能即将超过人类而忧心忡忡，而是担心AI将消解人类存在的意义，解构人类社会的关系。

2017年，AI领域的重磅论文《一种采用自注意力机制的深度学习模型Transformer》发表，这个不到200行代码的模型开启了AI发展的新阶段。原本分属不同领域的计算机视觉、语音识别、图像生成、自然语言处理等技术开始融合。在Transformer的模型下，工程师利用互联网上的文本进行AI训练，训练方法是在一句话里删除一些单词，让模型试着预测缺失的单词或接下来的单词。除了文本，此模型也可应用于声音和图像。和单词一样，工程师可以将声音和图像分解成细小的模块，并利用修补模型进行预测和填补。

“所谓生成式AI，通俗来说就是让AI能够像人类一样说话、写字、画画，甚至分析和理解问题。”张伟强说，基于这种“创作”能力，“人工”与“非人工”的边界正在消弭，数字世界的信息真伪也越来越难以辨识。目前，已有聊天机器人被用来生成针对性的网络钓鱼邮件。不久的将来，当人们听到或看到家人的声音或图像时，或许首先要问自己一个问题：这是真的吗？毕竟，最新的AI技术只需3秒就能拷贝一个人的特征。

此外，生成式AI还带来了其他新风险的挑战。张伟强举例说，第一个风险就是加深“信息茧房”。过去，当我们搜索信息时，还能得到多种答案以供选择。语言大模型则更像一个“茧房”，你将如同《楚门的世界》中的男主角，被动接受模型世界给你的信息。

大模型带来的第二个新风险是对创新动能的干扰。人类总是会在思考的过程中迸发灵感，在动手的过程中有所收获，在不断试错的过程中走向成功，不少伟大的发明都是研究的“副产品”。而大模型提供了前往正确答案的直通车，人们将由此减少很多试错机会。正如刘慈欣在科幻作品《镜子》中描写的一种人类“结局”，人类因为从不犯错而走向灭亡。

生成式AI如同不少伟大的发明一样，也会带来新的风险。图为电影《楚门的世界》剧照和刘慈欣的科幻作品《镜子》。

我们为AI“套笼头”的速度并不慢

生成式AI的飞速发展，让人类社会面临着一场信任危机。当网络上充斥着越来越多不知真假的图片和视频，AI助手几秒钟就“洗”出了一篇像模像样的稿件，大批学生开始用ChatGPT写作业、写论文，我们是否有信心用好生成式AI这个工具？

对此，梁正比较乐观。他认为，新技术总是伴随着风险，而人类曾无数次处理过这种情况，因此不必太过担心。在他看来，人类应对这一波生成式AI的速度算得上及时。

去年11月30日，OpenAI推出ChatGPT。今年3月，英国政府发布了第一份AI白皮书，概述了AI治理的五项原则。3月底，意大利个人数据保护局（DPA）宣布从即日起禁止使用ChatGPT，限制OpenAI处理意大利用户信息数据，同时对其隐私安全问题立案调查。随后，德国、法国、爱尔兰等国家也开始准备效仿意大利的做法，加强对ChatGPT的监管。

在生成式AI的立法方面，中国与欧盟基本同步。4月11日，国家互联网信息办公室发布《生成式人工智能服务管理办法（征求意见稿）》。梁正认为，《管理办法》从三方面给生成式AI的发展戴上了“笼头”：一是大模型的数据来源要可靠；二是对AI生成的内容应履行告知义务；三是一旦造成损害，相关责任方需要承担责任。他建议，对生成式AI实行分级分类管理。比如，对某些高风险领域应该谨慎或严格控制使用生成式AI，而对一般的办公娱乐场合，只要标注出AI生成内容即可。

与其焦虑，不如用技术规制技术

如果把生成式AI比作“矛”，那么检测其安全性的公司就是“盾”。目前，在全球范围内，“盾”公司的数量并不多。由清华大学人工智能研究院孵化的瑞莱智慧（Real AI）就是一家“盾”公司，他们负责检测内容是否由AI生成，以及给大模型的安全系统“挑刺”。

“人类需要保持辨识信息真伪的能力，只要能识别出哪些内容是AI生成的，并精准告知公众，这项技术也没有那么可怕。”张伟强说，目前他们研发了一套AI内容识别系统，在识别能力上处于国际领先。

比识别AI内容更重要的，是弥补第二代AI本身的安全缺陷。张伟强解释说，AI的“智力”提高后，需要视其为社会生活中的一位新参与者。但第二代AI本身的运算过程是个“黑箱”，相当于你无法看透这位新伙伴的所思所想（可解释性差），且他还极易被欺骗犯错（鲁棒性差）。至今在大模型中无法彻底解决的“幻觉”问题就是由此产生，即使数据来源准确可靠，但大模型仍可能会“一本正经地胡说八道”。

不可否认，ChatGPT开启了一场全球范围的大模型“军备竞赛”，大厂纷纷发布各自的大模型系统，不少小公司也推出了基于自身领域的“小模型”。张伟强表示，市场的充分竞争固然有利于行业快速发展，但其先天的安全不足同样需要引起重视。比如，上个月，瑞莱智慧仅通过添加少量对抗样本，就让Meta发布的史上首个图像分割模型SAM失灵，显示出大模型在安全性方面仍然任重道远。

梁正认为，未来，当人们回望现在所经历的这个阶段，会清晰认识到AI的工具属性。为了保证它永远只是工具，我们必须及时跟进它的动向，敏捷治理，就像历史上人类曾经一贯为之的那样。

对话曾毅：为何签名支持暂停巨型AI研发

■本报记者 沈湫莎

3月底，非营利组织未来生命研究所发表一封公开信，呼吁暂停研发比GPT-4更强大的人工智能（AI）系统至少6个月。公开信获得了千余名该领域专家、科技人员、产业高层的签名，其中包括中国科学家曾毅。

作为联合国教科文组织人工智能伦理特设专家组专家，曾毅也是中国科学院自动化研究所研究员、人工智能伦理与治理中心主任。当下的人工智能亟需解决哪些负面影响？身为AI技术和伦理专家，为何会在这封引发空前关注的公开信上签名？围绕这些话题，本报记者与曾毅进行了对话。

文汇报：您为何认为暂停6个月AI大模型研发是必要的？如何看待此次辛顿用离职发出的AI警告？

曾毅：生成式AI大模型的发展速度可以在几周内将潜在风险转变为实存风险，不负责任地研发和使用AI可能在短期内对社会造成显著而急迫的风险。

眼下，人工智能企业都在布局大模型，但很少有人说如何去防范风险。正因为我们对AI潜在的风险还没有完全准备好，但已经开始过早且过于激进地尝试，所以我觉得暂停6个月是必要的，先去解决一些潜在风险，先好好思考一下。当然，6个月只是一个倡议，是一个可以尝试的时段，能够取得一定效果肯定比不做准备要强得多。但是，6个月之内是否能够充分思考相关风险，并做到防范？我觉得这是不够的。

我认为，应对AI潜在威胁与应对气候变化危机并不存在重要性的差别，AI带给人类的挑战可能更紧迫。而且，AI技术每往前发展一步，人类对于伦理安全方面的考虑也需要前进一步。AI治理对其未来发展而言已经不是一道选择题，而是必答题。

文汇报：人类历史上有过很多次由新技术诞生而引发的危机时刻，您认为此次生成式AI给人类带来的影响和以往一样吗？

曾毅：生成式AI技术有能力合成虚假信息，极大降低了社会信任，该项技术的滥用、恶用，使得现今眼见、耳听都难以为实。这对于人与社会之间的关系提出了前所未有的挑战。从长远来看，生成式AI大模型等技术正在试图模糊人与AI之间的界限。

现阶段的AI只是看似智能的信息处理工具，并不具有真正的理解能力和真正意义的智能。它可以用来辅助人类决策，却不能代替人类决策，因为它并不是责任主体，也不具有生命。然而，诸多应用领域正在尝试用其替代人类，无法作为责任主体的AI工具被错误地赋予责任主体的责任与义务，这将对人类与社会发展产生深远而广泛的影响。

文汇报：您认为人类有可能始终掌控AI，让它永远只作为工具存在吗？要实现这一目标，人类有哪些手段？

曾毅：我们没有选择，对于作为工具的AI，我们必须竭力做好伦理安全框架、测试和负责任的部署，同时引导AI赋能可持续发展目标的实现。目前，AI赋能的实践主要在教育与健康领域，因为这两个领域有更明确的经济收益。而对于一些全球性的议题，例如生物多样性保护、气候变化、公平公正等对于人类的生存与发展极为重要的问题，AI的贡献微乎其微，这是人类使用AI这一工具时必须意识到并采取行动的。

我们应从制度建设和技术护航两方面引导AI向善。从制度建设与实践角度讲，制定伦理安全原则、规范及相关立法是必要的社会保障。从技术护航角度看，通过AI技术实现伦理安全关切，落地法律法规要求，并根据目前的技术瓶颈及时与制度建设互动，实现敏捷自适应治理是关键所在。

文汇报：在您的设想中，未来人类与AI应当是一种什么样的关系？

曾毅：现在，AI正以一种试探、尝试的方式去改变社会，人类社会则是以一种被动的姿态去应对，而不是我们所期待的主动去迎接技术变革，更多人是疲于应对。

如果把技术发展脉络放到更长的历史当中来看，人类应该回归其自身的位置，从事更多具有创造性的、情感交互的工作。比如在家照顾孩子和老人，关爱社会中的年长者和弱势群体，这些工作恰恰是AI最难替代的。在城市发展中，有必要给人留出没有必要使用AI的空间，并且让人们可以有选择权，决定使用或者不用这项技术。

文汇报：不少专家认为，AI最大的影响在于教育。在您看来，AI时代应培养孩子的哪些能力？

曾毅：现在有很多家长非常焦虑，总是在想我的孩子是不是应该尽早地接触人工智能。在我看来恰恰相反。如果真的学有余力的话，让孩子去学一门心理学，接触认知心理学的基础，让他对人本身有更多的认识——人类的智慧到底是怎么回事？我们的内心是什么样的？人如何为人？事实上，生成式AI让人更明白“何以为人”。因此，我觉得孩子们有必要学点哲学，在年轻的时候就去思考这些问题。