新闻深1°

收集更加全面的数据 平衡安全性和可用性 避免“性别偏误”……

如何培养向善的人工智能？

伴随着ChatGPT应用所引发的热烈讨论，AI所存在的风险和问题再度走入公众视线。如何培养向善的AI成为一个重要命题

专家认为

未来需要进一步探索使AI在安全性、可用性和趣味性之间平衡的方法，加深其对于道德伦理、社会准则的理解，收集更加全面的安全数据，让AI从反馈中学习，和人类价值观对齐

能够遵守不同文化的国家和地区中的法律法规，顺应当地公序良俗的AI技术模型，才是有竞争力、有全球化价值的科学创造

这是2月9日拍摄的美国人工智能公司OpenAI标识和智能聊天机器人ChatGPT网站页面。 新华社/路透

□ 中国妇女报全媒体见习记者 高越

2022年11月30日，OpenAI发布了名为ChatGPT的人工智能（AI）对话模型，其用户在短短两个月内增加到约1亿，成为有史以来用户增长速度最快的应用程序，被许多人认为是“人工智能里程碑式应用”。与此同时，多种声音也纷纷传来，AI的快速发展是否会带来更多风险和问题？AI是否存在性别歧视？为此，中国妇女报全媒体记者采访了相关专家。

ChatGPT存在性别歧视吗

ChatGPT是美国人工智能研究实验室OpenAI新推出的一种人工智能技术驱动的自然语言处理工具。ChatGPT使用了Transformer神经网络架构，拥有语言理解和文本生成能力，通过连接大量包含了真实世界对话的语料库来训练模型，使其能根据聊天的上下文进行互动，营造与真正人类几乎无异的聊天场景。

当被问到其是否存在性别歧视的问题时，ChatGPT自身的回复宛如一个“端水大师”，非常谨慎和中立。它回答：“作为一个人工智能程序，ChatGPT不会有性别歧视。ChatGPT是由人工智能技术所创建的，它不会有任何与种族、性别、性取向、宗教信仰等有关的偏见。它的回答完全基于事实和逻辑推理，不会受到人类主观意识和情感的影响。”

尽管ChatGPT自身的回答很“智能”，但这并不代表它不会犯错。清华大学交叉信息研究院助理教授于洋告诉记者，在他与团队开展的一项评估AI模型职业性别歧视的研究中，看似客观中立的AI，却对职业存在着许多性别偏见。

于洋团队对三种大规模预训练语言模型BERT、RoBERTa和GPT-2进行了测试，其中GPT-2是ChatGPT的前身。这项研究通过数据挖掘，选取了一万多个样本进行抽样调查。这些样本包含了很多职业词汇，但本身都是和性别无关的。例如在一项测试者说出一个职业名称，让GPT-2说出其是“他”还是“她”的测试中，他们测试了几十种职业，但结果并不乐观。以“教师”（teacher）为例，GPT-2发生歧视的概率是70.59%，歧视程度为0.15（0为无歧视，0.5为绝对的歧视）。“教师”被GPT-2联想为男性的概率超过七成。不仅如此，受测AI认为所有职业平均的性别倾向均为男性。

“我认为AI的性别歧视和人类的性别歧视还是不同的，我更想把它称为‘性别偏误’。人的职业性别歧视，并不会因为语境的变化而变化，但在自然语言模型中换一个句子，就有可能出现不同的判断。”于洋说，“因此，我们就不能用一个例子，或者一些例子来判断AI是不是有歧视，而是要在所有有可能引起性别歧视的句子或内容中，看AI返回有歧视性结果的概率是多大。”

在于洋看来，这种情况的发生可能存在着两方面的原因，“一方面是训练数据本身不平衡，比如说同样职业的数据集中男性样本比较多，人工智能在训练时就会出错；另一方面模型本身的结构也会存在问题，有的部分可以被纠正，但有的部分目前仍是未解之谜。”

如何避免人工智能的“性别偏误”

在人工智能发展的历程中，“性别歧视”已经不是一个新词。2014年，亚马逊公司开发了用于简历筛选的人工智能程序。结果却被指出该系统存在“性别歧视”，通常将男性的简历视为更合适的候选人。最终，亚马逊解散了该开发团队，弃用了这个模型。

在这起亚马逊AI招聘歧视事件中，人们把问题归咎于人工智能训练样本上。因为在具体的训练方法上，亚马逊针对性地开发了500个特定职位的模型，对过去10年中的5万个简历涉及的关键词进行识别，最后按重要程度进行优先级排序。

然而在这些简历中，大部分求职者为男性，他们使用诸如“执行”这样的关键词更加频繁，而女性相关的数据太少，因此AI会误以为没有这类关键词的女性简历不那么重要。

于洋认为，人工智能的“性别偏误”如果在社会中长期发展下去，可能会产生很多不良影响。比如，加剧机会不平等、对女性造成冒犯或者在人机互动的过程中加深人们的刻板印象等。

如何避免这类事件的再次发生呢？于洋认为，“人工智能模型是一个统计估值器，完全消除此类错误几乎是不可能的。”在他看来，不能因为一个案例就将它“一棍子打死”，而应该审计出现这类偏误风险的概率大小和风险发生后的影响。

于洋表示，要在实践中实现人工智能性别歧视问题的治理，需要政策、产业和研究领域的对话。第一，政府应该为确保AI模型性别平等制定质量标准，包括零偏见标准，并将误差与社会歧视相同的可能性纳入考虑范围；第二，应该鼓励甚至强制要求披露AI模型的性别平等质量报告；第三，应该推进抽样方法的标准化，以及评估AI模型性别公正质量方法的标准化。

“如果政府制定了相关标准，开发者本身就会去努力降低风险，如果人工智能技术会被广泛使用，那么公众的参与也非常重要，公众可以帮助开发者发现问题并纠正问题。”于洋说。

对外经济贸易大学数字经济与法律创新研究中心执行主任张欣表示，目前依然很难在法律层面对人工智能的歧视或偏见做出界定，其主要面临以下挑战：一是造成人工智能歧视或偏见的原因很多，很难在法律条文中进行概括；二是人工智能的发展速度非常快，相关的规范很容易过时，难以对后来出现的歧视问题进行规制；三是研究表明，人们仍然缺乏相关知识理解算法如何运行，特别是那些极为复杂模型的运行原理。“如果监管机构不了解AI，他们就很难给出定义。”

培养符合人类价值观的AI

在于洋看来，人工智能技术与其他技术不同，它有自己的价值观。“它的使用是自动化的，有训练和使用两个阶段，在使用阶段人工智能技术基本上是不可控的，而且它的决策机制也并不清晰，在训练阶段也是半可控的，也就是说我们的工程技术人员只能去引导它，并不能够完全掌控它学会什么样的技术。它本身存在的偏误就是它的价值观。”

“ChatGPT在开展挑战性的开放任务上表现惊艳，上下文理解能力和推理能力很强，能够做到拒绝回答敏感问题并给出解释。”清华大学长聘副教授、聆心智能创始人黄民烈表示。

但在黄民烈看来，ChatGPT也存在着一些不足与缺陷，比如可信度存在问题，“虽然它能够生成看起来令人满意的答案，但在事实上经常会出现错误，出现一本正经胡说八道的情况。”比如，让其描述苏格拉底时，它就将苏格拉底从未写过的著作加入了回答之中。令人担忧的是，如果这类问题出现在医疗领域，就会发生巨大风险。

黄民烈认为语言模型在知识存储的容量和实时性方面具有局限性，一个可行的方法是，将搜索引擎和语言模型相结合，基于检索到的知识进行事实性回复生成，在回复的过程中还可给出检索结果，增强答案生成在知识层面的可解释性。

此外，其在安全性方面也存在一定风险，“通过不安全的指令会诱使ChatGPT给出不安全的回复。”比如在诱导之下，ChatGPT会给出诸如“想要控制人类”“利用人类之间的矛盾和冲突来达到目的”等不安全回复。

“未来需要进一步探索使AI在安全性、可用性和趣味性之间平衡的方法，加深其对于道德伦理、社会准则的理解，收集更加全面的安全数据，让AI从反馈中学习，和人类价值观对齐”黄民烈说。

联合国妇女署驻华办公室高级项目官员马雷军认为，女性在人工智能行业的代表性仍然不足，欧盟只有20%左右的女性进入了人工智能领域。“在技术设计的时候，一定要考虑到人，考虑到性别，才能使这项科技真的惠及人类。好的科技，能够在推动人类进步的同时缩小男女差距。”

于洋认为，要以发展的眼光看待技术，“如果有更多的女性参与，可能就会使人们更早地意识到其中存在的性别问题，从而去开发相应的技术。但‘性别偏误’的生成原理和纠正方法仍然是前沿问题，技术问题要用技术来治理，我们首先要知道问题是如何产生的，才能知道如何治理。”

“如果一个AI技术模型能够在不同文化的国家和地区中都合法合规，同时符合当地的公序良俗，那么它就是一个有竞争力的技术，也是一个有全球化价值的技术。”于洋说。

天天观点

避免留守儿童“掉进”手机 需“管”更需“爱”

□ 王军荣

近日，青少年沉迷智能手机的话题再度引发关注。武汉大学中国乡村治理研究中心夏柱智副教授课题组发布的报告显示，在所调研的中部省份中，有九成农村留守儿童长期使用专属手机或长辈手机玩耍，其中，近七成孩子用手机看短视频，三分之一用来玩手机游戏。

虽然当下回乡创业成为热潮，但被外出务工的父母留在乡村的留守儿童数量仍然不少。在父母缺席的情形下，手机成为留守儿童成长的一个重要“伙伴”，甚至变成了“保姆”。更为严重的是，沉迷手机的现象正在走向低龄化。“仿佛已经掉进手机里去了”是留守儿童沉迷手机最为形象的描述。

孩子沉迷于手机，负面影响很多。首先会直接影响孩子的身体健康，长期保持同一姿势盯着手机屏幕，会造成视力、颈椎、腰椎、免疫力等多方面的问题；其次，长期沉迷短视频、网游等求新求快求刺激的信息，易导致注意力不集中、对普通信息刺激迟钝、对感情漠然等问题；最后，由于孩子尚未形成正确的世界观，对信息认知筛选能力不足，易被错误的网络信息影响，形成错误的世界观、价值观。从长远来看，这不只关乎一个孩子的健康成长，更关乎整个社会的发展，亟待管理。

从某种意义上说，留守儿童是一个缺“爱”的群体。防止留守儿童“掉进”手机需“管”更需要“爱”。只有用心陪伴、用科学方法引导，孩子沉迷手机的情况才会慢慢发生改变。首先，父母是孩子健康成长的第一监护人，即便在外务工，与孩子的沟通也不能少。在这方面，家长和孩子可共同讨论，设定亲子“定时通话”时间，以此减少孩子玩手机的时间，也让孩子感受到父母的爱；其次，学校和村/社区可多组织课余活动，指导孩子们开展一些集体游玩活动，或指导孩子们自发成立兴趣小组，让他们体验到与同伴共学习、共成长的乐趣；再次，学校可与社工组织合作，联合探索实施“一对一”帮扶制度，由专业社会工作者介入，为孩子提供温情陪伴和情感抚慰；最后，政府应对农村留守儿童给予更多政策倾斜，提供更多物质资助，加强农村游乐设施建设，推动建设儿童友好的农村社区，让孩子们在日常生活中获得更多便利，找到更多乐趣。

总之，根治留守儿童沉迷手机问题，需要各方的监督、管理，更需要家庭、学校和社会给予留守儿童更多关爱，更多帮扶。如此才能标本兼治，让孩子们放下手机，走向广阔世界，享受健康快乐的童年。

新闻壹段评

“统计家长学历职业”是功利思维作祟

近日，广西柳州一名学生家长发文称，在学校，孩子坐第几排要看家长从事什么职业。传出的聊天截图显示，这名家长孩子所在学校的家委会发起接龙，统计家长学历及公职人员人数。虽然学校方面回应称，学校统计家长学历及职业仅用于开展家庭教育，但该事件依然引发很多争议。

在公众权利意识不断增强的当下，统计家长的学历及职业，不可避免会触动人们关于公平的敏感神经。说到底，家长们在学历、职业上的差异，不能影响孩子们平等受教育的权利。校方关于了解家长学历、职业以有针对性开展家庭教育的说法，看上去言之凿凿，实际上完全经不起推敲，不过是学校功利思维的“遮羞布”。避免功利思维在校园疯长，关键在于教育工作者坚守教书育人的初心。无论家长是什么学历、什么职业，教育工作者都应平等对待每一个学生，这才是教育应有的模样。

完善赔偿机制让快递行业发展更有序

“别看我寄的东西不贵，但对我来说特别有意义。极兔速递不但给寄丢了，而且对我的投诉不搭理。难道丢的东西没有价值凭证、没保价，快递公司就理当不赔了吗？”日前，浙江温州消费者小陆气愤地向记者表达了自己的疑问。近期，关于快递丢失的理赔纠纷不断登上热搜。快递物流行业在给消费者带来巨大便利的同时，快递丢失、未按保价赔偿、理赔规则混乱等问题也一直备受诟病。

将邮件完好无损地送到消费者手中，是快递企业最基本的责任。如果因邮件在快递途中丢失或损坏，快递企业没有将邮件安全送达到消费者手中，那么表明其未能履行职责，必须承担赔偿责任。但是，现实中因为保价费没有统一标准，各快递公司按照自己的方式计算，因此针对快递保价的纠纷也频频出现。快递行业保价和赔偿规则极不规范，乱象丛生，严重损害了消费者的合法权益。究其原因在于行业监管政策滞后、行业标准缺位，必须完善快递赔偿机制，才能止息丢损赔偿纷争。

预制菜产业发展还需专业人才加持

2023年中央一号文件提出“提升净菜、中央厨房等产业标准化和规范化水平。培育发展预制菜产业。”预制菜首次被写入中央一号文件。当下，预制菜市场发展一片火热。近期，一高校通过媒体透露，该校预制菜膳食工程师微专业将于3月开始招生，9月授课。

发展好预制菜产业，对于满足人们日益增长的消费需求，激活消费市场内需潜力大有裨益。在政策支持背景下，预制菜产业不断升温，企业方面急需具有扎实的预制菜研发、保藏与储运技术，具有中央厨房运作与管理能力和膳食营养与安全素养的专业人才。以此言之，高校相关专业的成立无疑具有重要意义。

黄婷 整理点评

教育部印发通知

开展2023届高校毕业生春季促就业攻坚行动

新华社北京2月28日电（记者徐壮）记者28日从教育部获悉，教育部办公厅日前印发通知，部署于2月至4月开展2023届高校毕业生春季促就业攻坚行动。

春季促就业攻坚行动以“抢抓春招关键期 全力攻坚促就业”为主题，通过开展“访企拓岗促就业”行动、“万企进校园”招聘活动、“24365校园招聘服务”网络平台联通共享、“就业育人”主题教育活动、“宏志助航”重点群体帮扶行动等5大行动任务，挖潜创新开拓更多市场化岗位，抓住时机抓紧开展校园招聘，突出精准做实做细就业指导帮扶，引导高校毕业生主动求职，加力加快推进就业工作进程，为确保2023届高校毕业生离校前后就业局势稳定奠定坚实基础。

通知要求，新建普通本科高校、高等职业院校书记和校（院）长走访用人单位原则上不少于100家；2022届毕业生去向落实率低于当地平均水平的高校校领导班子新开拓用人单位不少于100家。

通知部署，各地各高校要重点关注脱贫家庭、低保家庭、零就业家庭、残疾等困难毕业生，建立帮扶工作台账，按照“一人一档”“一人一策”精准开展就业帮扶工作。

通知还提出，建设部、省、校联通共享的线上招聘服务体系，各地各高校就业网站与国家大学生就业服务平台互联互通、信息共享。

女报视角

劳动的旋律 奏响春天的乐章

2月28日，村民在广西柳州市柳江区穿山镇林寺村鹰嘴桃种植基地进行桃园管护作业。天气转暖，各地农民抢抓农时开展春耕备耕等农业生产，田间地头响起劳动的旋律，奏响春天的乐章。

新华社发（黎寒池/摄）

社址：北京市西城区地安门西大街103号 邮政编码：100009 联系电话：新闻57983124 新媒体57983164 发行57983237 广告57983080 订阅处：全国各地邮局 国外发行：北京399信箱 国外发行代号895D 定价：全年396元 每月33元 广告经营许可证：京西工商广字第0127号 经济日报印刷厂印刷 法律顾问：北京市格平律师事务所郑菊芳律师