

怎样应对人工智能带来的伦理问题

人工智能的持续进步和广泛应用带来的好处将是巨大的。但是，为了让人工智能真正有益于人类社会，我们也不能忽视人工智能背后的伦理问题。现在的人工智能界更多是工程师在参与，缺乏哲学、伦理学、法学等其他社会学科的参与，未来跨学科的人工智能伦理测试需要加强研究。

□ 曹建峰

人工智能时代加速到来，算法决策兴起

第三次 AI（人工智能，以下简称 AI）浪潮已经开启。在技术层面，有算法的进步。当 1956 年人工智能开始起步的时候，人们更多是在说人工智能；在第二次浪潮期间，机器学习成为主流；这一次则是深度学习，是能够自我学习、自我编程的学习算法，可以用来解决更复杂的任务。此外，计算能力的提升，包括现在的量子计算机，以及越来越普遍的大数据，对人工智能的作用和价值也非常大，使得更复杂的算法成为可能。在应用层面，从语音识别、机器翻译到医疗诊断、自动驾驶，AI 应用在不断加深、不断成熟，甚至已经开始超越人类，引发人们关于失业的担忧。同时也让人们开始期待具有通用智能的终极算法。在商业层面，面对可预期的好处和利益，国内外主流的互联网公司如腾讯、谷歌等都开始向 AI 看齐，AI 领域的创业和投资在如火如荼地进行着。全球已经有超过 1000 家 AI 公司，市场规模增长空间是非常大的，未来八年内将超过 350 亿美元。

在此背景下，各种互联网服务中越来越多地看到人工智能的影子，人们日益生活在算法之下，算法决策开始介入甚至主导越来越多的人类社会事务。比如，人们在互联网上获取的内容，诸如新闻、音乐、视频、广告等等，以及购买的商品，很多都是推荐引擎个性化推荐给用户的，而不是有人在背后决策。再比如，在金融领域，算法可以决定是否给某个用户发放贷款，以及具体的贷款额度。此外，一家美国投资公司早在几年前就开始研发管理公司的 AI 系统，招聘、投资、重大决策等公司事务都由这个 AI 系统来管理并决策。也许在未来，一家公司的成功不再主要依赖于拥有像乔布斯那样伟大的 CEO，而是一个足够智能足够强大的 AI 系统。更有甚者，英国、欧盟等都在大力推动人工智能技术在政务和民生方面的更深应用，政府服务不仅是数字化，而且是智能化。

人工智能伦理问题日益凸显

人工智能的持续进步和广泛应用带来的好处将是巨大的。但是，为了让 AI 真正有益于人类社会，我们也不能忽视 AI 背后的伦理问题。

第一个是算法歧视。可能人们会说，算法是一种数学表达，是很客观的，不像人类那样有各种偏见、情绪，容易受外部因素影响，怎么会产生歧视呢？之前的一些研究表明，法官在饿着肚子的时候，倾向于对犯罪人比较严厉，判刑也比较重，所以人们常

说，正义取决于法官有没有吃早餐。算法也正在带来类似的歧视问题。比如，一些图像识别软件之前还将黑人错误地标记为“黑猩猩”或者“猿猴”。此外，2016 年 3 月，微软公司在美国的 Twitter 上线的聊天机器人 Tay 在与网民互动过程中，成为了一个集性别歧视、种族歧视等于一身的“不良少女”。随着算法决策越来越多，类似的歧视也会越来越多。而且，算法歧视会带来危害。一方面，如果将算法应用在犯罪评估、信用贷款、雇佣评估等关切人身利益的场合，一旦产生歧视，必然危害个人权益。另一方面，深度学习是一个典型的“黑箱”算法，连设计者可能都不知道算法如何决策，要在系统中发现有没有存在歧视和歧视根源，在技术上是比较困难的。

为什么算法并不客观，可能暗藏歧视？算法决策在很多时候其实就是一种预测，用过去的的数据预测未来的趋势。算法模型和数据输入决定着预测的结果。因此，这两个要素也就成为算法歧视的主要来源。一方面，算法在本质上是“以数学方式或者计算机代码表达的意见”，包括其设计、目的、成功标准、数据使用等等都是设计者、开发者的主观选择，设计者和开发者可能将自己所怀抱的偏见嵌入算法系统。另一方面，数据的有效性、准确性，也会影响整个算法决策和预测的准确性。比如，数据是社会现实的反映，训练数据本身可能是歧视性的，用这样的数据训练出来的 AI 系统自然也会带上歧视的影子；再比如，数据可能是不正确、不完整或者过时的，带来所谓的“垃圾进，垃圾出”的现象；更进一步，如果一个 AI 系统依赖多数学习，自然不能兼容少数族裔的利益。此外，算法歧视可能是具有自我学习和适应能力的算法在交互过程中习得的，AI 系统在与现实世界交互过程中，可能没法区别什么是歧视，什么不是歧视。

更进一步，算法倾向于将歧视固化或者放大，使歧视自我长存于整个算法里面。算法决策是在用过去预测未来，而过去的歧视可能会在算法中得到巩固并在未来得到加强，因为错误的输入形成的错误输出作为反馈，进一步加深了错误。最终，算法决策不仅仅会将过去的歧视做法代码化，而且会创造自己的现实，形成一个“自我实现的歧视性反馈循环”。包括预测性警务、犯罪风险评估、信用评估等都存在类似问题。归根到底，算法决策其实缺乏对未来的想象力，而人类社会的进步需要这样的想象力。

第二个是隐私忧虑。很多 AI 系统，包括深度学习，都是大数据学习，需要大量的数据来训练学习算法。数据已经成了 AI 时代的“新石油”。这带来新的隐私忧虑。一方面，如果在深度学习过程中使用大量的敏感数据，这些数据可能会在后续被披露出去，对个人的隐私会产生影响。所以国外的 AI 研究人员已经在提倡如何在深度学习过程中保护个人隐私。另一方面，考虑到各种

服务之间大量交易数据，数据流动不断频繁，数据成为新的流通物，可能削弱个人对其个人数据的控制和管理。当然，现在已经有一些可以利用的工具来在 AI 时代加强隐私保护，诸如经规划的隐私、默认的隐私、个人数据管理工具、匿名化、假名化、差别化隐私、决策矩阵等等都是在不断发展和完善的一些标准，值得在深度学习和 AI 产品设计中提倡。

第三个是责任与安全。霍金、施密特等之前都警惕强人工智能或者超人工智能可能威胁人类生存。但在具体层面，AI 安全包括行为安全和人类控制。从阿西莫夫提出的机器人三定律到 2017 年阿西洛马会议提出的 23 条人工智能原则，AI 安全始终是人们关注的一个重点，美国、英国、欧盟等都在着力推进对自动驾驶汽车、智能机器人的安全监管。此外，安全往往与责任相伴。如果自动驾驶汽车、智能机器人造成人身、财产损失，谁来承担责任？如果按照现有的法律责任规则，因为系统是自主性很强的，它的开发者是难以预测的，包括黑箱的存在，很难解释事故的原因，未来可能会产生责任鸿沟。

第四个是机器人权利，即如何界定 AI 的人道主义待遇。随着自主智能机器人越来越强大，那么它们在人类社会到底应该扮演什么样的角色呢？自主智能机器人到底在法律上是什么？自然人？法人？动物？物？我们可以虐待、折磨或者杀死机器人吗？欧盟已经在考虑要不要赋予智能机器人“电子人”的法律人格，具有权利义务并对其行为负责。这个问题未来值得更多探讨。此外，越来越多的教育类、护理类、服务类的机器人在看护孩子、老人和病人，这些交互会对人的行为产生什么样的影响，需要得到进一步研究。



（上接 1 版）在学习了《习近平的七年知青岁月》采访实录后，我们认为这是一部非常生动、十分厚重的思想政治工作教材，要成为当前和今后很长一个时期我国高校开展思想政治教育的重要内容。一个个鲜活生动的案例充分体现了青年习近平强大的人格魅力和人格力量。这正是我们党对青年大学生的要求，是我们国家和民族对青年大学生的期待。

从我国长期的思想政治工作经验来看，树立榜样是最有效的方式。今日的中国比历史上任何时期都更接近中华民族伟大复兴的目标，比历史上任何时期都更有信心、更有能力实现这个目标。这样一个时代，将是中华民族历史上又一个辉煌的时代，将是世界历史上又一个震撼的时代。时势造英雄，英雄人物总是从时代风云中脱颖而出。复兴中华的奋斗，需要大批的民族精英。我们既需要在战场上勇敢杀敌的“战斗英雄”，也需要在科技兴国、科技强军的时代能够制造“两弹一星”的“科技英雄”。伟大的时代更需要伟大的人民领袖，这是我们国家和民族最宝贵也最重要的“大英雄”。当代青年学子已经不会经历那段知青下乡的

历史，也不会感受贫穷饥饿的折磨。但是，我们应该向不同时代的英杰们学习，从伟大人物的成长经历上汲取自身奋发向上的力量。当年七年知青岁月让习近平总书记深刻了解我们复杂的国情，了解我们这个国家和民族的根之所系、魂之所在，艰苦的生活锻造了习近平总书记钢铁般的担当精神和重于泰山的责任意识，塑造了他为民干事的人生观和价值观，为今后的伟大事业奠定了基石。习近平总书记身上散发出的伟大精神是一种大公大德，既体现了优秀青年的成长历程，又体现了人民领袖的精神特质。

“青年学子学青年习近平”学习教育，是引领青年学子成长成才的重要抓手，有助于青年学子深入了解习近平总书记成长为人民领袖的奋斗历程，深入学习青年习近平的优秀品质和高尚人格，树立正确的群众观，培养厚重的家国情怀，激发苦干实干、艰苦奋斗的昂扬斗志，在为祖国和人民建功立业中绽放青春光芒、健康成长成才。

学习时报：请问浙江外国语学院在开展“青年学子学青年习近平”学习教育过程中有何经验做法？

宣勇：在省委书记车俊同志的亲自关心

下，在省委宣传部、省委教育工委的大力支持和精心指导下，学校党委立足青年学生的思想实际，以高度的政治责任感推动学习教育在全校迅速铺开，扎实推进，全面深入，并取得了实效。

注重加强组织领导。学校主要领导亲自抓，班子成员自觉当好教育实践活动的组织者、推动者和参与者，带头学习、带头讨论、带头实践，切实发挥示范表率作用。通过专题党委会议研究部署，制定实施方案，针对师生群体的不同特点，设计丰富多彩、喜闻乐见、便于参与的活动载体；建立工作机制，由党委宣传部牵头负责，各职能部门通力合作，各学院深入开展，形成齐抓共管的工作格局。

注重营造浓厚氛围。学校在官网官微、校报校刊、广播台、宣传栏等宣传阵地开设“青年学子学青年习近平”专栏，推送活动新闻、学习体会、学习资料近 200 篇，运用官网、官微等宣传阵地，推送 45 期共计 77 篇学习资料。校报以“学、思、悟、践——青年学子学青年习近平”为题，进行 3 次专版报道，形成了浓厚的学习氛围。学校面向全部 281 个班级、2050 个寝室放发了 20 余万字的

构建算法治理的内外部约束机制

一是合伦理的 AI 设计，即要将人类社会的法律、道德等规范和价值嵌入 AI 系统。这主要是电气和电子工程师协会、英国等在提倡。可以分三步来实现。第一步是发现需要嵌入 AI 系统的规范和价值，存在道德过载和价值位阶的问题，即哪些价值优先，哪些价值在后。第二步是将所发现的规范和价值加入 AI 系统，需要方法论，有自上而下和自下而上两种路径。第三步是对已经嵌入 AI 系统的规范和价值进行评估，看其是否和人类社会的相一致。一方面是使用者评估，需要建立对 AI 的信任，比如当 AI 系统的行为超出预期时，要向用户解释为什么这么做。另一方面是主管部门、行业组织等第三方评估，需要界定价值一致性和相符性标准，以及 AI 可信标准。

但是需要解决两个困境。其一是伦理困境。比如，在来不及刹车的情况下，如果自动驾驶汽车往前开就会把三个闯红灯的人撞死，但如果转向就会碰到障碍物使车上的五个人死亡。此时，车辆应当如何选择？在面对类似电车困境的问题时，功利主义和绝对主义会给出不同的道德选择，这种冲突在人类社会都是没有解决的，在自动化的场景下也会遇到这样的问题。

其二是价值对接的问题。现在的很多机器人都是单目的的，扫地机器人就会一心一意地扫地，服务机器人就会一心一意给你去拿咖啡，诸如此类。但机器人的行为真的是我们人类想要的吗？这就产生了价值对接问题。就像 Midas 国王想要点石成金的技术，结果当他拥有这个法宝的时候，他碰到的所有东西包括食物都会变成金子，最后却被活活饿死。为什么呢？因为这个法宝并没有理解 Midas 国王的真正意图，那么机器人会不会给我们人类带来类似的情况呢？这个

问题值得深思。所以有人提出来兼容人类的 AI，包括三项原则，一是利他主义，即机器人的唯一目标是最大化人类价值的实现；二是不确定性，即机器人一开始不确定人类价值是什么；三是考虑人类，即人类行为提供了关于人类价值的信息，从而帮助机器人确定什么是人类所希望的价值。

二是在 AI 研发中贯彻伦理原则。一方面，针对 AI 研发活动，AI 研发人员需要遵守一些基本的伦理准则，包括有益性、不作恶、包容性的设计、多样性、透明性，以及隐私的保护，等等。另一方面，需要建立 AI 伦理审查制度，伦理审查应当是跨学科的，多样性的，对 AI 技术和产品的伦理影响进行评估并提出建议。

三是对算法进行必要的监管，避免算法作恶。现在的算法确实是越来越复杂，包括决策的影响都是越来越重大，未来可能需要对算法进行监管。可能的监管措施包括标准制定，涉及分类、性能标准、设计标准、责任标准等等；透明性方面，包括算法自身的代码透明性，以及算法决策透明性，国外现在已经有 OpenAI 等一些人工智能开源运动。此外，还有审批制度，比如对于自动驾驶汽车、智能机器人等采用的算法，未来可能需要监管部门进行事先审批，如果没有经过审批就不能向市场推出。

四是针对算法决策和歧视，以及造成的人身财产损害，需要提供法律救济。对于算法决策，一方面需要确保透明性，如果用自动化的手段进行决策决定，则需要告知用户，用户有知情权，并且在必要时需要向用户提供一定的解释；另一方面需要提供申诉的机制。对于机器人造成的人身财产损害，一方面，无辜的受害人应该得到救助；另一方面，对于自动驾驶汽车、智能机器人等带来的责任挑战，严格责任、差别化责任、强制保险和赔偿基金、智能机器人法律人格等都是可以考虑的救济措施。

在今天这个人工智能快速发展，人类在诸如围棋、图像识别、语音识别等等领域开始落后于人工智能的时代，对人工智能进行伦理测试同样重要，包括道德代码、隐私、正义、有益性、安全、责任等等，都是十分重要的。现在的 AI 界更多是工程师在参与，缺乏哲学、伦理学、法学等其他社会学科的参与，未来这样跨学科的 AI 伦理测试需要加强研究。因为在某种意义上我们已经不是在制造一个被动的简单工具，而是在设计像人一样具有感知、认知、决策等能力的事物，你可以称其为“更复杂的工具”，但不可否认，我们需要确保这样的复杂工具进入人类社会以后和人类的价值规范及需求相一致。

践，引导和激励广大青年学生学思结合，知行合一。在暑期期间，学校将组织 180 余支社会实践队伍、1000 余名师生赴梁家河和浙江各地，追寻习近平总书记足迹，将社会实践活动进一步做实做细。

注重坚持长效机制。学校把学习教育与推进“两学一做”常态化制度化有机融合，把习近平总书记青年时代的经历作为中心组学习、党员培养、支部工作、兵团学习、校园文化、社会实践等的重要内容，持续组织好师生学习；把学习教育与课堂教学有机融合，把青年习近平的优秀案例融入课堂教学中，着力推进“思政课程”向“课程思政”的转变；把学习教育与立德树人相结合，渗透到学校教育育人的各个环节，着力推动形成全员、全程、全方位育人的工作格局。

接下来，我们将抓好“青年学子学青年习近平”学习教育的持续和深化工作，争取取得更大成效。

学习时报：预祝这项学习教育向深度和广度推进，创造出新的经验。我们将持续跟踪这项学习教育的进展情况并适时加以报道。

宣勇：谢谢《学习时报》。