算法“黑箱”下人工智能安全存疑

可解释性让AI更透明

□ 科普时报记者 陈 杰

“AI的应用落地在不断提速，但在算法‘黑箱’下，AI的不可解释性让‘黑箱’很难释明，进而让人们对AI的安全难言放心。”日前，在2022年科技向善创新周上，与会专家围绕AI的透明和可解释难题从技术角度直指产业痛点。

与此同时，由腾讯研究院、优图实验室等多家机构组成的跨学科研究团队历时一年完成的《可解释AI发展报告2022》也引发热议。报告从可解释AI的概念、监管趋势、行业实践、发展建议等热点问题出发，抽丝剥茧地呈现可解释AI产业的现状和发展趋势，以期能为产业解决AI“黑箱”释明难题提供一份参考和借鉴。

可解释AI已成热点

目前，部分人工智能应用已成为通用型的技术，而人类对AI则一直有更高的期待。不少人坚信，科幻电影《Her》中的AI机器人可对行为作出解释，从而帮助人类作出决策，甚至与人类产生深度交流的一幕，迟早也会出现在人们的日常生活中。

虽然产业的发展已经起步并快速成为行业热点，但这些过于理想化的场景目前还只能存在于科幻作品中，实践中可解释AI不论在技术上还是体验上都还没达标。

如今，金融机构的贷款审批都是基于AI作出决策，如果这一复杂的算法风控系统拒绝申请，那么贷款人就没办法获得贷款。这只是AI在众多日常场景应用中的一幕，只要是AI做出了决策，公众就必须接受结果，至于什么原因，算法没有也不可能给出解释。

“有时候连开发人员都不能很好地理解AI算法‘黑箱’运作的具体细节，这就导致了AI模型的透明度和可解释性的问题。”腾讯研究院秘书长张钦坤表示，如果不解决这两个问题，不仅影响用户对AI应用的信任，而且也可能会带来算法歧视、算法安全和算法责任等方面的相关问题。

其实，人工智能的可解释以及科技伦理等问题已经成为AI领域的必选项，2021年起始的“十四五”规划里面也明确强调要健全科技伦理的体系。

厦门大学人文学院院长朱菁认为，AI可解释性问题之所以受到重视，主要因为AI的发展虽然变得越来越强大，但也变得越来越“黑”，再者AI虽然变得越来越实用，其可靠性和应用范围也得到提高。“在一些重要的应用领域，对于AI可信任性、安全性等方面的要求也越来越高，可解释性就是满足上述要求的认知基础。”

目前来看，国内企业在可解释AI实践方面还比较零碎，没有达到系统化的程度，但就整体而言，产业一直都是朝着可解释的方向发展。

鱼和熊掌不可兼得

AI快速深入日常生活，既带来了兴奋和期待，也带来一些忧虑，甚至恐慌。AI到底是人类的好帮手，还是会成为人类强劲的对手？其实这很大程度上取决于人类对AI的理解。只有打开了人工智能的黑箱，了解到AI背后的机制，并认为它是可解释、可理解的，人类在这种共识下才能真正达成对AI的信任。

然而，当前的AI可解释性及透明度与算法的高效率还是一个矛盾体。

微众银行首席人工智能官杨强认为，AI算法高效率的同时，可解释性就很差。同样，AI线性模型的准确率没有那么高，它的可解释性相对就会强一些。“这就意味着我们要在可解释和高效率两个维度上做一个取舍，但目前并没有两个维度都高的AI算法。目前可解释AI在各个行业都是一个起步，也都不可或缺，但如何做好AI可解释的量化，才是当前业界该考虑的重点。”

“不同人群、不同应用场景对AI算法的可解释性期待是不一样的，不应该搞一刀切。现阶段深度学习普遍缺乏可解释性的情况下，AI透明度就尤其重要。”腾讯天衍实验室负责人郑冶枫表示，AI算法应尽可能详尽披露模型、训练数据分布等情况。

朱菁也认为，目前来说，AI的可解释性和预测的准确性二者不可兼得，既提高预测准确性，同时还要把可解释性提高，要实现这件两全其美的事情难度比较大。其次，解释的多元化除了怎么都行的这种形态以外，还有孤岛式的、互动式的以及整合性的形态。“这些只是同一个整体的不同侧面，AI能不能做到这些，现在其实并不是很清楚。”

推动可解释模型构建

可解释是可信AI的重要组成部分，也是可信的前提条件之一，它有很强的独特性。当前可解释AI研究思路很多，但并没有一个明确的框架，毕竟解释对象的不同，框架也难统一。

香港中文大学（深圳）吴保元副教授认为，与其说AI的可解释性，还不如称之为AI的可解释力。“可解释性可能会让人们误认为这是一种性质，而可解释力就是一种可解释的能力，就如同人们常说的理解力、领导力，是一种手段，一种行为，一种操作的存在。”

《可解释AI发展报告2022》从科技向善的角度出发，认为需要找到一个平衡的可解释AI路径来实现可信AI，确保科技向善。张钦坤解释说，就是在设计可解释性要求时，需要充分考虑可解释性要求和其他重要的伦理价值和目的，可解释本身不是目的而是手段。“在设计可解释AI的时候，首先需要考虑实现什么样的目标，其次才是思考在特定的情境下如何更好地匹配这些目标。”

对于AI研究里的解释性问题的基本立场，朱菁的建议是解释的多元化：“针对不同的问题，哪怕是在同一个领域里，也不能指望只有唯一的解释方式或模式，要允许多种多样的解释模型存在。”

除了“五感”，人体可谓“百感交集”

□王 欣

明代著名思想家王阳明有句名言：“你未看此花时，此花与汝同归于寂；你来看此花时，则此花颜色一时明白起来。”从生理学的角度理解就是：如果一个事物你完全感觉不到，对于你而言就等于不存在。双目失明的人无法感受春天万紫千红的美丽，双耳失聪的人无法领略音乐动人心弦的魅力。感觉是心灵与世界沟通的桥梁，是人生宝贵的财富。

感觉如此重要，人类到底有多少种感觉？许多人会脱口而出：视觉、听觉、味觉、嗅觉、触觉。“五感”的确是人们非常熟悉也重要的感觉，但是感觉的种类绝不止这五种，它们种类繁多得超乎人们的想象。

视觉、听觉、味觉、嗅觉这四种头面部的感觉合称特殊感觉，对应眼睛、耳朵、舌头、鼻子这四种器官。它们“特殊”在哪里？就在于感觉器官的结构复杂、功能强大。就拿视觉来说，它接收的信号是光波，光波进入眼的折光系统，聚焦在视网膜上。视网膜对光波的强弱、波长等参数进行编码，通过视神经传入脑，各级视中枢核团传到大脑的视皮层，于是看见了远处的风景。

感觉是真实的吗？《心经》有句名言：“色不异空，空不异色；色即是空，空即是色。”颜色、声音、气味、味道这些感觉都不是真实存在的，是被大脑“创造”出来的，真实存在的是光波、声波、化学分子等客观物质，感觉是头脑对这些客观物质的主观反映。如果我们置身于外太空，会发现那里寂静无声，不是声波消失了，也不是耳朵失灵了，而是声波在真空中无法传导，可见感觉必须依靠外界刺激和内在机能相互作用才会浮现。

“五感”的触觉在生理学上称作躯体感觉，它的感受器广泛分布于皮肤的真皮层和皮下组织。触觉感受器包括毛囊感受器、麦斯纳氏小体、鲁菲尼氏小体和帕西尼氏小体，分别感受轻触觉、触压觉和振动觉。头面部的触觉由脑神经传入大脑皮层，躯干和四肢的触觉由脊髓上行到脑。

“五感”没有提到痛觉和温度觉。痛觉包括来自皮肤的体表痛，肌肉、肌腱和关节的深部痛和内脏器官的内脏痛，它相当于人体的报警装置，保护着我们免受伤害，提醒我们及时就医；另一方面，剧烈而持续的疼痛本身也是疾病。痛觉感受器多为游离神经末梢，其上分布着各种受体和离子通道，可以被各种伤害性刺激激活。温度觉感受器在体表呈点状分布，分为冷点和温点，冷点比温点多。它们遍布全身皮肤，于面部、手背、前臂掌侧面、足背、胸部、腹部以及生殖器的皮肤较密集。

“五感”也没有提到平衡觉和本体感觉。平衡觉感受器是内耳的前庭器官，主要功能是感受机体姿势、运动状态以及头部的空间位置。晕车就是因为头部的晃动使前庭器官过度亢奋，不断发出“头部位置不正”的信号，其主观感受就是眩晕。本体感觉的感受器是肌梭与腱器官，它们位于肌肉和肌腱，感受肌肉张力与长度的变化并调节肌肉张力，这使我们在运动的时候肢体伸缩自如、力量恰到好处。平衡觉和本体感觉通常传递到脑干和小脑，进而引起姿势反射，不抵达大脑皮层，主观上不被感知。

人体中为数众多的内脏感觉更是不为人知。以血压的调节为例，血压升高时，血管壁内的压力感受器发出的冲动就会增加，传入延髓的心血管中枢，通过自主神经减慢心率、舒张血管、降低血压。已发现的内脏感觉有数十种，加上前面提到的各种感觉，人体真可谓是“百感交集”。

（作者系华中师范大学副教授、湖北省生理学会理事）

冬奥会火炬“飞扬”的科技“外衣”

□ 科 文

1月5日，冬奥会短道速滑冠军张会在活动现场手持火炬进行展示。

当日，北京2022年冬奥会火炬展示活动在黑龙江省哈尔滨市举行。本场活动结束后，冬奥火炬还将陆续在大庆、齐齐哈尔进行展示。 新华社记者 张涛 摄

2008年奥运会上的祥云火炬因为造型独特优美，给人们留下美好的记忆，而即将到来的冬奥会上的火炬，也有一个响亮的名字——“飞扬”。

在上海石化厂区的车间内，工作人员们正忙着组装北京冬奥会和冬残奥会火炬“飞扬”。

2021年2月4日，在北京冬奥会开幕倒计时一周年活动上，火炬“飞扬”揭开面纱。从表面看，火炬“飞扬”外形极具动感和活力。在设计上，为了衬托北京即将成为奥运历史上首座“双奥之城”，“飞扬”的外观与北京2008年奥运会开幕式主火炬塔形态相呼应，以祥云纹样“打底”，自下而上，从祥云纹逐渐过渡到雪花图案，最后在顶端化身为“飞扬”的火焰。

不仅有漂亮的外观，“飞扬”的外壳也蕴含着“黑科技”。中国石化上海石化公司副总经理黄翔宇表示，火炬的点火系统全部包在外壳里面，从外面是看不到的。火炬的外壳采用了碳纤维材料，手感非常轻。

中国石化上海石化创新研究院总经理林生兵表示，碳纤维的质量只有钢的1/4左右，但是强度是钢的7至9倍。这次研发团队用碳纤维与树脂形成的复合材料来做奥运火炬，堪称世界首创。

“冬奥组委给它的第一个评价就是轻，而且很牢固，随便怎么摔也摔不坏，另外一个就是冬季火炬传递的时候天气很冷，复合材料解决了这个问题，避免了冬季传递火炬手感冰凉。”黄翔宇说。

由于北京冬奥会火炬接力将在冬季低温环境中进行，“飞扬”采用氢做燃料，除了氢具有环保的特点，还因为氢燃料的特性保证了火炬能在极寒天气中使用。

黄翔宇说，火炬的火是从里面烧出来的。一般的复合材料都不能在火里烧，正好上海金山区有一家企业研发了一种树脂，这个树脂是耐火的，经过测试分析，最后通过工艺调整达到了既能够耐温又能够耐火的要求，800℃、900℃都可以。

负责生产树脂的企业负责人刘章友表示，作为第三代树脂材料的聚硅氮烷树脂，兼具有机物附着力强大与无机物耐高温的特点，同时集耐腐蚀、磨损和防污防水、超薄膜等优势于一身，恰好能解决火炬所需的各种要求。

在生产车间里，由石油产品加工成的这一条条黑色丝束，每一束都包含着12000根碳纤维丝，再经过三维立体编织最终做成的火炬外壳，看不出任何接缝与孔隙，整个造型浑然一体。

科技冬奥伴我行

异种器官移植，未来还有多远的路要走

（上接第1版）

2017年，《自然》杂志发表的一篇论文向世人公开了一项激动人心的研究结果：美国哈佛大学杨璐菡教授和她的团队通过内源性基因编辑的方法，利用基因剪刀技术完成了猪内源性逆转录病毒相关酶基因的敲除，并成功克隆出无内源性病毒的“猪2.0”。

此外，杨璐菡还通过另外的实验证明，携带内源性病毒的猪不会将病毒传染给胚胎，无内源性病毒感染的小猪在成长的过程中不会重新被病毒感染，人的内源性病毒逆转录酶也不能引起猪内源性病毒的继续复制。

“尽管异种器官移植研究取得了很大进展，但移植器官的功能兼容性问题却成了尚待解决的技术挑战，被移植的猪器官能否完全发挥原有人体器官维持荷尔蒙分泌、代谢平衡等功能仍有待检验。”杨璐菡说，解决供体猪的异种病毒传播风险和免疫兼容性只是“万里长征第一步”，目前还需要不停地摸索和改进异种移植器官在功能上的兼容性。

潘登科认为，即使异种器官移植的技术难题得以攻破，但在未来的临床应用上，异种器官移植仍面临伦理、监管等诸多挑战。

中国企业在异种器官移植领域蓄势待发

世界卫生组织鼓励各国开展异种器官移植的临床研究，并指出该类研究必须防范潜在风险，相关产品需在严格监督下生产，实行高标准质控，对异种移植研究者也提出了非常严格的要求。作为前沿科技，异种器官移植赛道的成员屈指可数，目前仅有美国、德国、中国和韩国等少数国家，而且大部分以实验室研究为主。

虽然头部领跑企业主要为美国企业，但中国企业也在蓄势待发。目前国内进行异种器官移植研究的企业主要有赛诺生物和中科奥格等。

2005年8月9日国际在线报道，中国第一头自主完成的克隆猪诞生了，其中一名主要的科研人员就是潘登科。2010年11月，中国农业科学院北京畜牧兽医研究所潘登科科研团队成功培育出4头适宜人体器官移植的基因改造猪。潘登科说，这4头猪是我国首次消除超级性免疫排斥基因的一种器官移植猪。2018年，中国农业科学院将10余年科研成果——基因修饰猪研发的核心技术及基因编辑猪种群，转让到中科奥格，将实验室技术转化为产业技术。

潘登科介绍，我国在基因编辑猪到非人灵长类肾脏移植研究方面，采用人体临床免疫抑制方案，深入开展了异种器官移植的免疫基础研究。

针对全球首例基因修饰异种心脏移植到活人体内的手术，潘登科说，目前异种器官移植主要是在没有其他方法可用的情况下进行的试验治疗。随着我国在该领域的快速发展，关于异种器官移植的审批、监管和伦理制度等都亟待建立。

火山喷发会减缓气候变暖吗

（上接第1版）

“人造火山喷发”无法解决气候变化

在汤加火山喷发事件发生后，中国社会科学院生态文明研究所研究员，IPCC第五、第六次评估报告第三工作组主要作者陈迎一直对此保持关注。她注意到社交媒体上一个引起热烈讨论的话题：“如果火山喷发有降温作用，那我们是不是只要人工制造这种气溶胶，并将其播撒到大气平流层中，就不用花大力气减排了？”

对此观点，陈迎表示反对，“如果没有减排这个前提，只靠太阳辐射干预（SRM），即通过人为方法大尺度改变地球系统的辐射平衡以应对全球变暖，肯定是无法解决气候变化问题的。”

陈迎补充说，SRM也解决不了海洋酸化问题，同时还可能带来其他风险和不确定性，比如改变气温和降水分布等。可以确定的是，SRM无法作为应对气候变化的“主力”。不过，根据近年最新研究，如果建立在大幅度减排基础上，SRM有潜力作为应对气候变化的辅助措施。

浙江大学地球科学学院大气科学系教授、IPCC第六次评估报告第一工作组主要作者曹龙介绍，目前提出的SRM方法主要包括向平流层注入气溶胶、海洋低云亮化、增加海洋和陆地表面的反照率。这些方法的基本出发点是增加地—气系统的反照率，减少到达大气和地面的太阳辐射，通过短波辐射干预的方法，抵消温室气体增加造成的暖化效应，但无法在全球和区域尺度上，完全抵消温室气体增加引起的气候变化，并且SRM无法缓解海洋酸化。

那么，从模拟走向现实，SRM还需要多久？曹龙坦言，由于目前对云—气溶胶辐射过程的相互作用和微物理过程认知仍很有限，对于基于气溶胶的SRM冷却潜力认知还有很大的不确定性，并且IPCC第六次评估报告第一工作组报告对于SRM气候效应的评估主要集中在全球尺度，缺乏针对SRM对不同区域气候影响的评估。在下一步研究工作中，有必要利用包括更完备的云—气溶—辐射过程的高分辨率模式，对SRM方法进行模拟研究，进一步认知不同SRM方法的冷却潜力和对气候系统的影响。另外还要大力加强在不同地点和时间实施的不同SRM方法，对全球和区域气候影响的研究。

健康码的“安全”也需护航

□ 科普时报记者 陈 杰

1月17日，一年一度的“春运”正式启动，除“例行”的抢票热点话题之外，已经启动全国一码通行的“健康码”也成为公众关注的焦点。

近期，国内个别地方健康码先后出现系统问题，不少网友便担心“春运期间健康码的亮码频次暴增，会不会出现迟滞影响出行”。在疫情持续这两年里，健康码亮码已超数百亿次，先后有10余个城市曾发生过系统故障。

基于此，国家发展改革委、交通运输部、公安部等14个部门日前发布的《关于全力做好2022年春运工作的意见》中提出，要落实健康码全国“一码通行”，鼓励各地区推广刷身份证自动核验健康码服务。各地健康码业务系统也未雨绸缪，全力保证春运期间业务安全平稳运行。

日前，一款可对健康码业务系统运行提供保障的产品“一码盾”上线。该产品由信息安全等级保护关键技术国家工程实验室携手知道创宇公司联合推出，可远程布防且只需15分钟即可生效，能提供一站式压力缓解、应用加速、流量管控、抗DDoS（分布式拒绝服务）、安全DNS（域名系统）、防黑客攻击等功能，可使健康码业务系统负载显著降低。

“二维码业务系统是近两年来被广泛部署的系统，并没有经过充分的压力验证，在面临突发情况时容易产生稳定性问题，造成业务中断。”知道创宇CTO杨冀龙在接受记者采访时表示，作为当前疫情防控的最重要的工具，健康码具有7×24小时高可靠、高可用的要求，特别是在疫情突发的关键时刻，其使用频率可能会有数百倍甚至数千倍的激增，短时间内数千万到亿次级的亮码需求；再加上健康码承载公民敏感个人隐私数据，其安全防护更是必不可少，公民的个人隐私数据必然会成为网络黑客的攻击目标，对平台建设者而言是个极大的挑战。

杨冀龙形容健康码业务系统就像一个高速收费站，正常情况下收费站可轻松应对车流，但面对节假日等流量暴增的情况，收费站就特别容易阻塞瘫痪，这时就需要加派工作人员带着卡到收费站外进行分发，提高收费站通过效率。“其实，在互联网上也有CDN（内容分发网络）的标准型产品，专门负责‘外援’工作，可大大提高‘发卡’效率。一码盾便是基于创宇盾SCDN（安全域名系统）在全国拥有的50个核心机房、超3000个安全内容分发节点、6T骨干网络带宽储备，可应对上千亿级业务请求，从而有效保障健康码业务的连续性。”

从实际部署情况来看，一码盾的接入方式非常简单，只需要一码盾服务将域名通过DNS的方式指向一码盾即可完成部署。当用户要访问健康码时，流量就会先到达一码盾部署在全国各地的数千台服务器上，一码盾平台在进行恶意请求过滤数据缓存之后，只有极少数的请求会被转发到客户的业务服务器来保障数据的准确性。

作为一款免费产品，一码盾的整个部署过程只需要几分钟即可完成，完全能应对春运期间跨省市的人员高流动的严峻考验。