在世人眼中一贯“高冷范儿”惯了的英国上议院开了一个新奇的“脑洞”：4月16日，英国上议院提交了一份长达183页的报告《人工智能在英国：充分准备、意愿积极、能力爆棚？》，提出在发展和应用人工智能过程中有必要把伦理道德放在核心位置，以确保这项技术更好地造福人类。这份报告提出之后，在全球范围内引发了舆论的关注：很多人担忧，我们真能把人类道德教给人工智能吗——尤其是在人类自己还不太搞得懂道德是什么的情况下。

面对人工智能 人类“有理”说得清吗

本报记者　王昱

人类对AI提出“新要求”

如果你是个科幻小说迷，那你一定知道著名科幻作家阿西莫夫提出的“机器人三定律”。1950年，在计算机还十分简陋的年代，阿西莫夫已经在小说《我，机器人》中表达了人类有一天会被自己制造的机器人奴役的忧虑，并给出了相应的解决办法：教会机器人讲道德。这就是所谓阿西莫夫三定律：第一、机器人不得伤害人类，或看到人类受到伤害而袖手旁观；第二、在不违反第一定律的前提下，机器人必须绝对服从人类给予的任何命令；第三、在不违反第一定律和第二定律的前提下，机器人必须尽力保护自己。

长久以来，阿西莫夫三定律一直被认为是科幻为科学作出的一大重要贡献，“机器人三定律”被称为“现代机器人学的基石”。言外之意，大多数人都认同，如果有一天我们真的能造出比人类更聪明、更强力的人工智能，就要按着阿西莫夫的意思给它讲道理。然而，在英国上议院新近公布的这份报告中，我们却发现，这个被推崇惯了的阿西莫夫三定律似乎不再适用了。

该报告提出，应确立一个适用于不同领域的“人工智能准则”，其中主要包括五项准则：人工智能应为人类共同利益服务；人工智能应遵循可理解性和公平性原则；人工智能不应用于削弱个人、家庭乃至社区的数据权利或隐私；所有公民都应有权利接受相关教育，以便能在精神、情感和经济上适应人工智能发展；人工智能绝不应被赋予任何伤害、毁灭或欺骗人类的自主能力。

很显然，准则从三条上升为五条，人类对人工智能的要求正在加码。这背后是否包含了公众对人工智能更深的忧虑，眼下还不好说。但可以肯定的是，很多人对这个新的“五准则”似乎并不满意。英国媒体就批评说：与逻辑严谨、权重明确的“阿西莫夫三定律”相比，新提出的“五准则”逻辑松散，彼此之间没有先后权重分配。交给人工智能去执行时，人工智能将会不知道应该优先保障哪条——比如说，当人工智能遭遇某种情况，为了“服务人类共同利益”需要侵犯少数人隐私权时，它们究竟该怎么做？新提出的“人工智能准则”并没有给出明确答案。但这种情况显然是多见的。

不过，虽然批评声音很多，英国上议院的这份报告还是不乏进步之处。比如，它明确点出了“人工智能应遵循可理解性和公平性原则”，就获得了相关研究者和舆论的一致点赞。但讽刺的是，这两点却最可能沦为空话。

人工智能愈发“不可理喻”

先说“可理解性”原则。

这一点其实非常重要，因为随着人工智能的飞速发展。越来越多的人开始意识到，仅仅像阿西莫夫三定律那般要求人工智能对人类“好”是不够的，人工智能还必须向人类说明，为什么它这样做是对人类“好”。否则人类将会陷入被人工智能强行操控的“奴役感”当中无法自拔，这就是所谓的“可理解性”原则。但遗憾的是，以目前的技术条件，“可理解性”原则恰恰难以达到。

我们并不完全清楚人脑的学习机制，讽刺的是，我们现在对人工智能如何思维也知之甚少，这通常被称为“黑匣子问题”——你知道输入的数据，也知道得出的结果，但不知道眼前的盒子是怎么得出结论的。而目前的人工神经网络尚不具备自我解说功能，即无法说明其所作所为是基于什么样的“数据”算出来的。这就导致了人工智能的所有决定的理由只能靠人类自己去“猜”，一旦猜不透，“可理解性”原则就无法达成。

这方面其实已经有很多例子，前不久，微软高级研究员卡鲁阿纳开发了一套人工智能系统，将医疗数据输入人工神经网络，包括症状及其后果，从而计算在任何一天患者的死亡风险有多大，让医生能够采取预防措施。效果似乎不错，直到有天晚上一位美国匹兹堡大学的研究生发现了问题。他用一个更简便的算法处理同一组数据，逐条研究神经网络做诊断的逻辑，而其中一条诊断令人匪夷所思：“如果你患有肺炎，那么患哮喘对你是有好处的”。

卡鲁阿纳说：“我们去问医生，他们说‘这太糟糕了，你们需要修正’”。哮喘是引发肺炎的重要风险因素，因为二者都会影响肺部。人们永远也不知道这个智能机器如何得出了哮喘对肺炎有益的结论。有种解释是，有哮喘病史的患者一开始患肺炎，就会尽快去看医生，这可能人为地提高了他们的存活率，因此人工智能就错误地认为有哮喘对肺炎是有帮助的结论。

如果这种猜测是正确的，那么人工智能做出的这个判断显然就是个错误结论。但问题是，我们也不知道人工智能是否就是这么想的，更不知道它利用相同思路算出了多少似是而非的结果。所以，结论是在人工智能无法掌握“自我解说”能力以前，“可理解性”原则永不可能实现。而想要让人工智能像人类一样解释自己的思维路径，天啊，那太难了——事实上，很多时候，人类自己都不知道某种结论是怎样在脑中形成的。

怎么让人工智能有道德

我们再来说说“公平性”原则，与“可理解性”原则在技术发展的前提下还有希望解决不同。目前，学界对“公平性”原则总体预期更为悲观——基于人工智能现有算法，“公平性”原则几乎不可能达成。

去年2月，美国普林斯顿大学曾捅了一个不小的娄子，该学校的某研究小组开发了一款具有自学能力的人工智能系统GloVe，该系统能够识别并且理解网络文字，并对文字给出“拟人”的情感表达，比如“鲜花”是令人喜悦的词汇，而“蜘蛛”则是令人不悦的词汇。在系统初步完成后，研究人员将GloVe放到网上去进行“深度学习”。但最终的学习结果令人大吃一惊：当研究人员将一串名字输入到GloVe系统时，惊奇地发现这个机器人居然学会了“种族歧视”。该系统将白人常用名字识别为“令人喜悦”的词，而非裔美国人最常用的一些名字却被划为“令人不悦”的词！由于该结果在美国十分政治不正确，吓得研究人员赶紧叫停了该实验。

人工智能是怎么学会“种族歧视”的？这其实很好解释。与人类一样，人工智能是通过“数学加权”（数学计算中将参数比重加入计算称之为加权）对数据进行演算的，如果人工智能在收集到的信息中，不断得到某个人名与“犯罪”“凶杀”“贫穷”等词汇一同出现，那么它就将利用参数将该人名“加权”为令人不悦的词，反之亦然。众所周知，在美国目前的社会中，黑人犯罪率确实高于白人，这不难出现歧视少数族裔的计算结果。令人深感讽刺的是，人工智能得出该结论的路径与人类的歧视思维其实如出一辙。若说有什么区别，那就只有人工智能因分析效率更高，“歧视效率”也更高，光听个人名就能歧视某些人。

更令人恐惧的是，如果按照该路径走下去，在未来，你可能不知干了什么就被人工智能“鄙视”了。所以若不改变加权算法，人工智能对人类的公平性将荡然无存。

其实，人类对人工智能的道德新要求之所以难以达成，说到底还是我们自己的问题：人类对很多自己的道德准则、思维方式尚且了解不深，却要求人工智能将它们的决策原理明白地解释给我们听；人类对很多同类的偏见就是基于既往经验的“数学加权”，却要求人工智能脱离这种偏见。这种拔着自己头发上天的思维显然是不现实的，道德有亏的人类不可能造出一个全知全善的人工智能。

因此，未来，我们究竟该怎样教会人工智能道德，这个问题恐怕还有很长的路要走。

喝醉酒说外语更溜？

发表在《精神药理学杂志》上的一篇论文显示，人们喝下低剂量的酒精饮品后，在外语口语上的表现更加流利，即使他们本人并不这样认为。

这项研究的对象是50名在荷兰马斯特里赫特大学学习的德国人，所有人都承认自己平时至少喝一点酒。为了能听懂用荷兰语讲授的课程，他们都在近期通过了一项初级语言能力测试，也就是说他们的荷兰语能力基本相当（都是菜鸟）。

在研究中，来自利物浦大学的心理学家克斯伯根要求每位被测试者与一名面试官用荷兰语进行五分钟的日常会话。在参与对话之前，他们随机地喝下水或者酒精饮料，其中的酒精含量依照一定比例根据被测试者体重算出。

所有谈话录音由母语为荷兰语的考官评分，而考官也无从得知每位被测试者是不是喝了酒。同时，被测试者也会根据评分标准和自身表现打一个自评分。结果显示，喝了酒的那部分人自信心并没有提高，他们的自评分跟没喝酒的那些人基本无异。但在评分考官看来，他们的表现的确更优，主要表现在流利程度，尤其是发音准确度上，其他诸如语法、词汇和逻辑方面则与喝水组持平。看来，语言能力的提高是客观事实，而不是个别酒鬼借醉胡侃一番之后出于错觉的自吹自擂。

该研究使用的酒精剂量非常低，毕竟一瓶啤酒远远不能把人灌醉，对于大多数人，你根本没法分辨出他到底喝了还是没喝。但研究人员指出，更高的剂量也许会起到反作用，因为喝太多只会麻痹舌头、让人口齿不清，而且损害认知、注意力、控制力和短期记忆功能。由于被测试者清楚地知道自己喝的是水还是啤酒，所以很难说他们表现更佳到底是因为酒精的生理作用，还是出于一种心理暗示。

当然人们还想知道，酒精是不是只对初学荷兰语的德国人有效。至少还有一项其他研究支持这个结论：发表于前些年的一篇论文显示，低剂量的酒精能改善美国人的泰国语发音。

虽然该研究并没有考量被测试者的精神或情绪状态，但论文共同作者、来自荷兰的科学家维特曼认为，这种现象的出现跟酒精帮人们克服了拘谨、压抑和紧张情绪有关。如果是在社交场合，一点酒精还能提升自信、缓解社交焦虑。这项研究给我们的启示大概在于，要想流利地说一门外语，除了勤学多练之外，提高胆量和放开自我也是非常重要的。

（据《南都周刊》）

发火有生物钟 黄昏焦躁症有望获治

美国科学家进行的一项小鼠实验发现，生物钟参与了调控动物的“进攻性”行为。这一研究结果有望用于治疗阿尔茨海默病患者的黄昏焦躁症状。

日前发表在英国《自然·神经学》杂志上的研究表明，雄性小鼠间为保护领地而发生的好斗行为在一天中的强度和频率会随光照发生变化。论文作者、哈佛大学医学院克利福德·塞珀教授说，小鼠在黄昏时最好斗，在清晨时最温顺，这表明进攻性行为接受光照后逐渐增强，在黄昏达到顶峰。

研究人员操纵负责调控生物钟的神经元，发现让某一神经递质失去活性后，小鼠进攻性行为的昼夜起伏就会消失，而这些小鼠总体会更好斗，进攻行为显著增加。

患有阿尔茨海默病和其他老年痴呆症的患者常出现“日落症候”，即在白天结束时突发性暴躁。塞珀认为，控制生物钟有可能会让患者更加平和。

（据《北京日报》）