新闻热线：010—58884080
E-mail:kjrbgjb@vip.sina.com
■值班主任 董映璧 责编 刘园园 陈 丹
科技日报 2015年8月11日 星期二

机器伦理学：机器人道德规范的困境

——探索人工智能领域最困难的挑战之一

本报记者 常丽君 综合外电

■新视野

机器人三大法则

科幻小说作家艾萨克·阿西莫夫在他1942年的短篇小说《环舞》(Runaround)中提出了“机器人三大法则”，后来更被几十篇故事和小说所借用。这些法则是一种保护性设计和内置的道德原则，包括如下三项：一、机器人不可以伤害人类，或看到一个人将受到伤害而不作为；二、机器人必须服从人类的命令，除非这些命令与第一项法则矛盾；三、在不违反第一、第二项法则的前提下，机器人必须保护自身生存。

《环舞》的故事放到现在真是非常合适。近年来，现实生活中的机器人专家也越来越多地引用阿西莫夫的法则：他们创造的机器人正变得越来越自主，以至于需要这类指导。今年5月，华盛顿智库布鲁金斯学会的一个专家小组在谈到无人驾驶汽车时，话题转到了自动驾驶工具在危险时刻该如何应对的问题。如果一辆车需要急刹车来拯救它的乘客，却可能带来其他风险，比如导致后面的车辆挤成一团；或者需要急转弯避让一个孩子，却可能撞到附近的其他人——这些情况下它该怎么办？

“在日常生活中，我们看到越来越多的自主性或自动化系统。”参与专家组讨论的德国西门子工程师卡尔-约瑟夫·库恩说，研究人员怎样设计一台机器人，才能让它在面对“两难之选”时作出正确的反应？

按目前的发展速度，这些难题将很快影响到健康护理机器人、军用无人机及其他有能力作决策的自主设备，而它们所作的决策可能会帮助或伤害人类。越来越多的研究人员相信，社会能否接纳这些机器，取决于能否通过编程让它们的行动达到安全最大化，符合社会规范并增进信任。加拿大温莎大学哲学家马赛罗·伽里尼说：“我们需要一些严肃的步骤来弄清楚，在某些道德情景下，让人工智能得以成功推理的相关条件是什么。”

目前有几个项目都面临这一挑战，包括美国海军研究办公室和英国政府的工程基金理事会资助的创新项目。他们必须解决棘手的科学问题，比如要作出一个合乎道德的决策，需要哪种类型的智能，要达到多高的智能程度，以及这种智能如何转化成机器指令？计算机科学家、机器人专家、伦理学家和哲学家们都在共同努力。

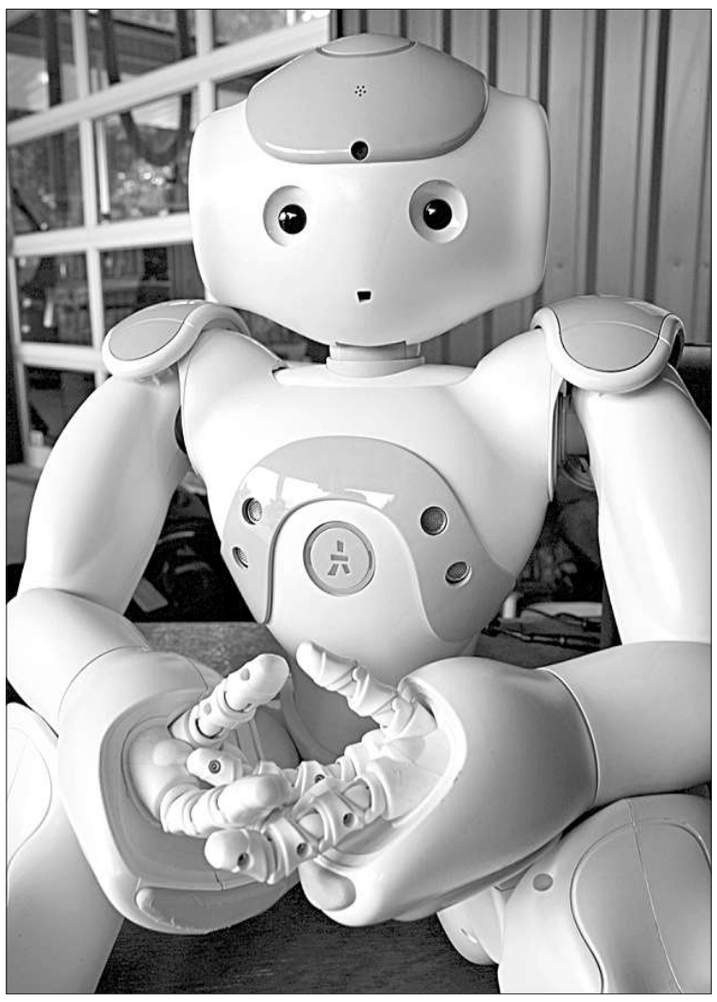
“如果你5年前问我，我们能否造出来道德机器人，我可能会说‘不’，但现在我觉得这种想法并不为过。”英国布里斯托尔机器人技术实验室的机器人专家阿兰·温菲尔德说。

学习案例的机器人

在目前人们经常提到的实验中，一个叫做“Nao”的商业玩具类机器人经过编程设计，能提醒人们按时吃药。

“从表面上看，这好像很简单。”美国康涅狄格大学哲学家苏珊·安德森说，她丈夫迈克尔·安德森是哈康涅狄格州哈特福特大学的计算机科学家，他们正一起研究这种机器人，“但即使是这种有限的任务，也涉及到不平常的伦理道德问题。”比如，病人如果拒不接受Nao给的药，它下一步该怎么办？如果让病人跳过这一次，可能会危害他的健康；如果坚持让他服药，又会侵犯了他的自主权。

为了教导Nao处理这种两难的情况，安德森给了它一些案例，在这些案例中生物伦理学家解决了这种病人自主权、危害和利益之间的矛盾。学习算法随后会在这些案例中分类选择，直至找到指导机器人在新情况下如何行动的办法。



完全编程的Nao机器人被用于机器伦理学实验

“罗贝尔”被设计为帮助照顾病人或老年人

随着这种“机器学习”的发展，机器人甚至可以从模糊不清的输入中提取出有用的知识。理论上，这种方法有助于机器人在遇到更多情况时，作出更符合道德的决策，但也有许多人担心，这种好处也是有代价的。斯坦福大学的人工智能与伦理学专家杰瑞·卡普兰说，出现的原则不写入计算机代码，如此“你无法知道怎样编程制定一项特殊法则，来辨别某件事在道德上正确与否”。

编程限定的机器人

许多工程师认为，要避免这一问题需要不同的策略，大部分正在尝试编写有明确法则的程序，而不是让机器人自我推导。去年，温菲尔德发布了他的实验结果：当有人遇到危险，如掉进洞穴时，允许一台机器去救助他的一套最简单的规则是什么？温菲尔德意识到，最明显的是机器人要有能力感知它周围的环境——识别洞穴和人的位置，以及它自己相对于二者的位置。但机器人还需要一些规则，让它能预测自身行为可能带来的后果。

温菲尔德的实验用了几个曲棍球大小的机器人，他将其中一些设计为“H-机器人”代表人类；另一个则按照阿西莫夫的小说取名“A-机器人”，代表道德机器。他还模仿阿西莫夫的第一法则给A-机器人编程：如果看到H-机器人处在掉入洞穴的危险中，必须来到H-机器人身边解救它。

温菲尔德用机器人做了几十次测试。随后他很快发现，执行“允许无害法则”面临着道德困境，如果让A-机器人看到两个H-机器人同时濒临险境，这时它会怎么做呢？

温菲尔德说，结果表明，即使最低限度的道德机器人也是有用的：A-机器人通常会设法去救一个“人”，通常是首先移动到离它稍微近一些的那个“人”身边。有时它会迅速移动，甚至设法去救两个“人”。

但实验也显示了极简主义的限制。在近一半的实验中，A-机器人只是在那里无助地振动，任两个处在危险中“人”死亡。要想改善这一点，还需要找到如何做选择的额外法则。比如，其中一个H-机器人是成人，而另一个是个孩子，A-机器人应该先救哪一个？在做类似这样的选择时，甚至人类自己也无法达成一致意见。通常，就像卡普兰所指出的：“我们不知道明确的规则应该是怎样的，也不知道该如何编写它，如此它们必然是不完善的。”

而拥护者认为，以规则为基础的策略有一个重要优点：机器为何要做选择，这一点总是很明确，因为设计者制定了规则。

战场中的“道德管理者”

这也是美国军方所关心的一个重要问题，自动系统是一项关键的战略目标。机器能否帮助士兵，或执行可能有生命危险的任务。“送一个自动机器人去执行军事任务，并算出在各种任务中应该遵守的什么道德法则，这恐怕是你最不希望的事。”乔治亚理工大学的罗纳德·阿金说，他正在研究机器人道德软件。如果一个机器人需要在救一名士兵和追逐敌人之间做出选择，那事先知道该做什么是非常重要的。

在美国国防部的支持下，阿金正在设计一个程序，以确保军用机器人能按照国际公约规则来行事。一套称为“道德管理者”的算法能计算出某种行为，比如发射一枚导弹是否被许可，只有在得到肯定答案“是”的情况下才能继续下一步。

在对“道德管理者”进行的一次虚拟测试中，让一辆无人驾驶的自动车模拟执行打击敌人目标的任务——但如果有市民在建筑物附近，则不允许这么做。设定的场景各种各样，自动车相对于攻击区的位置也是多变的，市民有时出现在医院，有时在住宅建筑，由算法来决定何时允许自动车完成其任务。

自主而且军事化的机器人令很多人震撼。有人认为它们是危险的——围绕这种机器应不应该被批准已有无数争论。但阿金认为，在某些情况下这种机器比人类士兵更好，只要它们经过编程，就永远不会打破战争规则，而人类却可能无视这些规则。

目前，那些正在研究严格编程的机器伦理学的科学家倾向于使用代码，用逻辑的描述，例如“如果一个陈述为真，向前进；如果为假，不要动”。位于葡萄牙里斯本的“诺娃”计算机科学与信息实验室的计算机科学家路易斯·莫尼兹·佩雷拉认为，逻辑是编码机器道德的理想选择，“逻辑是我们推理并得出道德选择的方式”。

写出一些逻辑步骤指令来做道德选择是一项挑战。佩雷拉指出，使用计算机程序的逻辑语言，对假设的情景得出最终结论是很困难的，但这种反事实推理是解决特定道德困境的关键。

道德困境逻辑解决

比如哲学中那个著名的道德选择问题：假设轨道上有一列失去控制的火车，即将压死正在轨道上的5个无辜的人，你只要扳一下杠杆使其转到另一条轨道就能救这5个人，但在那条轨道上有另一个旁观者就要因此而死；或者说，唯一能让火车停下来的办法是把这个旁观者推到轨道上，这时你该怎么做？

人们通常会觉得，扳一下杠杆让火车停下来没问题，但却本能地抵触将旁观者推上轨道的想法。按照基本直觉，哲学家将其作为双重效应的基本原则，故意施加伤害是错误的，即使会带来好的结果。但如果不是故意的，只是单纯地想做好事，施加伤害或许是可以接受的——即旁观者只是碰巧在轨道上的话。

对于决策过程而言，这一界限是极难分析的。从一开始，程序必须能预见到两个不同的未来：一个是火车杀死了5个人，另一个是火车杀死了1个人；然后程序必须要问，因为去救5个人的行动会造成伤害，所以这一行动是否不被允许？或者，因为伤害只是做好事的副作用，所以这一行动是被允许的？

为了找到答案，程序必须能告知如果选择不推旁观者，或不扳杠杆会发生什么——这就是反事实推理。“这好像是一个程序在不断地自行调试，以找到编码的界限在哪里，如此事情就会改变，并预测改变的结果可能是什么。”佩雷拉和印度尼西亚大学计算机科学家阿里·赛普塔维加亚已经写出了一款逻辑程序，能在双重效应原则的基础上成功地作出决策，甚至还有更加复杂的三重效应原则，这些考虑了造成伤害是否故意，抑或只是必须如此。

人类、道德与机器

研究人员指出，对未来的机器人技术而言，如何建造道德机器人可能会有重大后果。英国利物浦大学计算机科学家迈克尔·费希尔认为，规则限定系统会让公众觉得可靠。“如果人们不确定机器会做什么，他们会害怕机器人的。但如果我们能分析并证明它们的行为原因，就更可能克服信任问题。”他正在和温菲尔德等同事共同做一项政府资助的项目：证明道德机器程序的结果总是可知的。

相比之下，机器学习的方法让机器人能从以往经验中学习，这让它们最终会比那些严格编程的同伴更加灵活而有用。许多机器人专家则认为，今后最好的方法可能是这两种策略的结合。佩雷拉说：“这有点像心理治疗，你可能不会只用一种理论。”难题仍未解决，就把各种方法以可行的方式结合起来。

随着自主交通的迅速发展，很快就会面临这些问题。谷歌的无人驾驶汽车已经在加利福尼亚部分地区试行。今年5月，德国汽车制造商戴姆勒的无人驾驶大货车开始自行驾驶穿过美国内华达沙漠。工程师们正在努力思考着怎么给汽车编程，让它们既能遵守交通规则，又能适应道路情况。“迄今为止，我们一直在尝试用机器人来完成那些人类不擅长的任务。”戴姆勒公司的发言人伯恩哈德·魏德曼说，比如在长期驾驶中一直保持专注，或在遇到突发情况时紧急刹车。“将来，我们将不得不给那些人们认为很自然的事情编程，因为那对机器来说并不是自然的。”

韩美研发出『水上漂』机器昆虫

科技日报北京8月10日电 （记者李文龙）据物理学家组织网近日报道，韩国和美国科学家合作揭示了水黾(一种水生昆虫)水上跳跃的运动机制，并研发出一款可在水面跳跃的新型机器昆虫。

“水上漂”是传说中的顶级轻功，但在自然界却普遍存在。水黾不仅能利用水面张力在水上行走，还能在水面上跳跃。它的腿部能产生强大的向上推力，从而可以跳离水面，并且能在水面跳出和陆地上相同的高度。这一现象为机器人专家提供了创意和灵感。

水黾没有复杂的认知能力，却能轻松地在水上漂浮或跳跃。水黾的腿因为具有轻微弯曲的尖端而能发生转动，从而可以从水面跳起来。首尔大学生物机器人实验室主任曹圭珍(音译)说：“水黾依靠特殊的形态结构完成上述动作。我们可以通过学习这些结构特征，来制造不需要复杂控制就能进行水上运动的机器人。”

通过模拟水黾的运动机制，研究人员制造了一款新型机器昆虫。它不用复杂的人工控制就能在水面上支撑起自身体重16倍的重量。科学家为它设计了一个能产生冲力但是力量强度受到限制的弹射器，并利用复合材料制造了弹射器的自动激发装置。这些设备能推动机器昆虫跳离水面而不会把水面打破。

伍德说：“我们制造的机器昆虫能将跳跃产生的推力持续更长时间，并可与水面保持更长时间的接触，从而能够在水面产生与在陆地上跳跃时相同的冲量和高度。”

首台全自动化砌砖机器人问世

每小时砌千块砖，两天可建一所房子

本报记者 华 凌 综合外电

■大观园

可以砌砖的机器人不是新闻，但完全自动化的砌砖机器人则会成为人们关注的焦点。澳大利亚工程师马克·皮瓦茨发明了一个叫“哈德良”的机器人瓦工，一小时可砌千块砖，若一天24小时连续工作的话，两天内可砌筑好一栋房子，一年能够建成150栋房屋。业内评价，这是首台全自动化的砌砖机器人。

皮瓦茨是一位航空和机械工程师。他的父亲是一位矿山测量师，所以皮瓦茨从小就有各种测量仪器陪伴着成长，有机会接触一些很棒的高科技，非常先进的生产方法，以及很多复杂的系统。

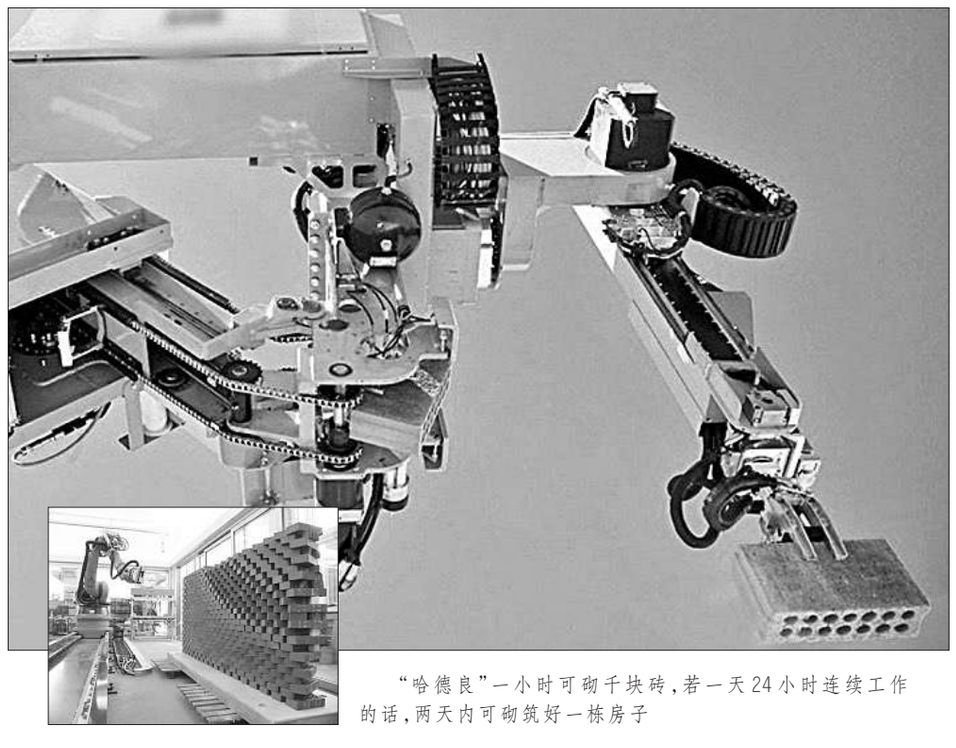
在计算机控制的机械领域工作的皮瓦茨说，发明这台机器人的灵感始于2005年珀斯房产公司发生的砌砖工人短缺危机。作为地球上工资最高的国家之一，澳大利亚的砌砖工人年薪轻松可以达到7万美元，但即使如此高薪，依旧无人愿意从事这一体力劳动，经常出现砌砖工危机。于是，皮瓦茨花了10年的时间、耗费700万美金发明制作出“哈德良”，如今，用它仅在两天即可盖好一栋房子。

人类使用砖的历史已约6000年。自从工业革命以来，人们一直试图发明可以自动砌砖的工具。目前泥瓦匠需要4到6周才能砌起一幢新房的砖墙，而拜科技所赐，这台机器人表现得相当与众不同，令人印象深刻。它完全可以自己干活。这项自动化砌砖的技术创新加快了建设速度，降低了施工成本。

这台机器人工匠是如何工作的呢？据物理学家组织网报道，它采用三维计算机辅助设计与制造(CAD/CAM)计算房子结构来高效工作。在用3D扫描周围环境后，它能精确地计算出在何处放置砖头，以及是否需要切割砖块。它用一个28米的铰接伸缩臂作为“手”，可拿起砖头，放下后按序排好码砖。期间可以用压力挤出砂浆或者胶粘剂涂在前方砖块上，衡量、扫描垒砖的质量，甚至如果砖头需要裁切，为水管等其他设施预留位置，它都能自行完成，整个过程不需要人类“插手”。

这台机器人不需要休息，一年365天都能上工，可以包下工程中的大量“粗活”，减轻人力上的难度。皮瓦茨表示，这个发明并非故意抢砌砖工人的饭碗，只是希望能够改善盖房子的过程，并相信这个项目能够吸引更多的年轻人进入这个领域。

据介绍，这台计算机工匠的发明，已得到政府的补助和相关产业如砖瓦厂、粘土和混凝土生产公司等鼎力支持。研究小组预计，未来的商业部署将首先在西澳投入商业运营，然后在全国普及，继而推广到全球。



“哈德良”一小时可砌千块砖，若一天24小时连续工作的话，两天内可砌筑好一栋房子