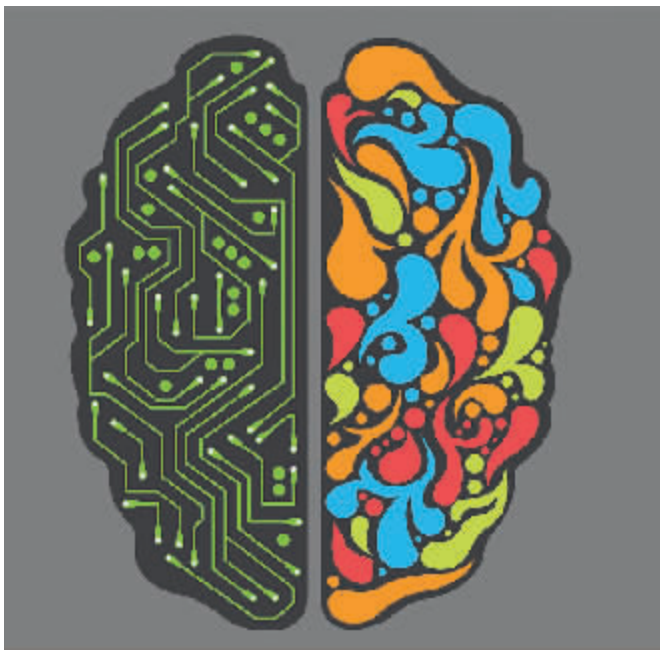
（上接 8 版）

出过这个想法。但最近人们在人工智能方面取得的进展又引起了新的担忧，牛津大学哲学家尼克·博斯特罗姆（Nick Bostrom）就提出了“回形针最多化”（paperclip maximiser）的思想实验。这个实验假设人工智能希望能收集尽可能多的回形针。它会想尽一切办法来收集回形针，并且通过自我升级来找到收集回形针的新方法，而且还会反抗一切阻止它这样做的企图。最后它“把整个宇宙空间都变成了一个回形针制造工厂”。正如博斯特罗姆所说的：“人工智能不需要人类一样的行为和心理动机。它们可能不会出现人类常会犯的错误，但是会犯别的错误，例如执着于收集回形针。它们的目标一开始可能看起来是无害的，但如果它们能够自主升级自己的性能就会非常危险。即使是运行在一台不联网的计算机上的被束缚的超级人工智能也会竭尽全力地劝说它的主人让它获得自由。因此，超级人工智能不仅仅是一种新技术，而是对人类的威胁。”技术专家尽管不相信人工智能会失去控制，但普遍还是会有社会伦理上的忧虑。要避免“弗兰肯斯坦回形针”隐忧，必须尽快建构起相应的伦理框架和法律规制。

一是合伦理的人工智能设计，并在人工智能研发中贯彻伦理原则，即将人类社会的法律、道德等规范和价值嵌入人工智能系统。这主要是电气和电子工程师协会（IEEE）、美国、英国等在大力提倡。

早在上世纪 40 年代，科幻小说家阿西莫夫就提出了著名的机器人三定律：（1）机器人不得伤害人类个体，或者目睹人类个体将遭受危险而袖手不管；（2）机器人必须服从人给予它的命令，当该命令与第一定律冲突时例外；（3）机器人在不违反第一、第二定律情况下要尽可能保护自己的生存。也有学者提出兼容人类的人工智能三原则：（1）利他主义，即机器人的唯一目标是最大化人类价值的实现；（2）不确定性，即机器人一开始不确定人类价值是什么；（3）考量人类，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期望的价值。



2017 年 1 月在美国加州召开的“阿西洛马会议”，近千名人工智能和机器人领域的专家们联合签署了称为“阿西洛马人工智能原则”，呼吁全世界的人工智能工作遵守这些原则，共同保障人类未来的利益和安全。“阿西洛马人工智能原则”的核心是“为了人类的人工智能”。在“阿西洛马人工智能原则”23 条中，其中具有纲领性的条目有：第 1 条“研究目的”确立了人工智能研究的目的是创造服务于人、并为之所控的人工智能和机器人原则，这个原则是人机之间的基本伦理保证，而这个保证最先由研究和开发人员的伦理意识体现在研发中并遵守；所以第 16 条“人类规制”表达了一种热切的期望。第 2 条“研究经费”中对人类资源、意志、价值体系的基础地位的坚守。第 10 条“价值归属”、第 11 条“人类价值观”，进一步提升为机器对人的价值归属是高级人工智能的设计准则；第 9 条“责任”则将人机伦理关系作为责任明确地加在研究人员身上；第 12 条“个人隐私”和第 13 条“自由和隐私”体现了人的尊严。第 6 条“安全性”是为研究工作设定机器对人的保障条件。第 7 条“故障透明性”、第 8 条“司法透明性”，是对人工智能现有研究、开发的法律规制方面的伦理要求。第 14 条“分享利益”、第 15 条“共同繁荣”、第 17 条“非颠覆”、第 18 条“人工智能军备竞赛”，则是人类社会的伦理原则在人工智能研究领域的具体体现。上述原则都是基于人的目的而理解的人–机关系，“阿西洛马人工智能原则”第 23 条“共公利益”甚至包括了“超级智能”，但这仍是作为全人类整体利益的考虑。为了人类的人工智能，这种把尊重人的道德地位放在人工智能领域不同层次上多次被强调，都是基于人的道德和人类尊严的朴素性。

人工智能的发展对人类自身的地位和价值产生了影响，对此担忧由来已久。其中由一些著名的人工智能问题专家以志愿者身份创立的“未来生命研究所”，于 2015 年 1 月发表了一封公开信《为稳健性和利益性的人工智能而进行研究》提出，人工智能必须只做我们要它们做的，即“有益的人工智能”口号。同样，此前也已经引起了业内外的广泛注意和严肃讨论并有所行动，如：IEEE 关于人工智能及自主系统的伦理考虑的全球倡议，人工智能未来总统委员会，人工智能百年研究计划，人工智能伦理与治理基金会，艾伦人工智能研究所，等等。

二是必须对人工智能应用进行必要的监管，避免作恶，同时针对人工智能可能带来的风险及造成的人身财产损害，提供必要的法律救济。现在的人工智能越来越复杂，决策的影响越来越大，未来可能需要对其进行监管。责任、透明性、可审核性，并使无辜受害者不至于无助地沮丧和恐惧。可能的监管措施包括标准制定，涉及分类标准、设计标准、责任分担等等；透明性方面，包括人工智能技术代码和智能决策的透明性，如果用自动化的手段进行决策，则需要告知用户，保障用户知情权，并在必要时向用户提供解释，国外已经有 Open AI 等一些人工智能开源运动。此外，还需要确立审批制度，比如对于无人驾驶汽车、智能机器人等试验和应用，未来需要监管部门进行预先审批，未经审批向市场推出必须审慎。对于人工智能造成的人身财产损害，无辜的受害者应该得到及时合理的救助；对于无人驾驶汽车、智能机器人等带来的挑战，厘清责任分担、差别化责任、强制保险、确立智能机器人法律人格等都是可以考虑的法律救济措施。

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研究日益重要，包括道德代码、隐私、正义、有益性、安全、责任等等。遗憾的是，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。未来人工智能伦理需要加强跨学科的研究，我们的研究必须直面这些伦理道德挑战，毕竟它关乎着我们人类的未来。

（作者为上海师范大学马克思主义学院副教授）

学人在读

贾敏（中国浦东干部学院教师、历史学博士）

最近读完且颇感具有分量的著作，要数知名华裔学者、美国布鲁金斯学会约翰·桑顿中国研究中心主任李成老师的新书《观念的力量：崛起中的中国智库与思想者》（World Scientific，2017）。

众所周知，智库研究是当今中国社会科学界至为关心的议题之一，国家从顶层设计层面也出台了一系列助力智库发展的重要文件，而从学术发展自身的规律而言，寻求学术话语与现实问题相结合，智库建设确实提供了一个殊为难得的广阔平台，为走出象牙塔创造了条件。

不必讳言，在经历过去几年智库建设与发展的狂飙突进后，不少人都在思考智库繁荣现象背后出现的各类问题，特别是有学者提出的中国智库建设过程中的“跟风造库”、“库多智少”、“有库无智”等突出现象后，有关智库建设的利与弊成为业内同行们绕不开的话题。

在本书作者看来，当前中国智库跨越式发展所折射出来的，恰是中国改革发展进入了新阶段，因此，可以从更为广阔的政治经济与社会文化背景中寻求中国智库崛起的答案。李成近些年来关注中国社会各阶层中的新兴力量与思想潮流，以及知识分子、学者讨论并参与公共政策背后的观念逻辑，提出了一系列有价值的观点。进而，这些内容需要有效传递给中国以外的读者与研究者，从而避免让他们对中国当前的智库发展产生千篇一律、或是运动式兴起的刻板印象。

书中对于当前中国智库发展过程中的一些突出现象和问题，都做了非常坦率的解读。由于中美两国社会发展阶段的不同，一些美国特有的现象，如智库“旋转门”机制，并不一定适合中国；中国智库发展过程中如何使自身定位清晰化，如何处理各方面关系，也是智库长期发展与健康运行的前提保证。

本书内容布局兼具宏观性的整体介绍和具体智库的个案介绍，也包含作者对于若干具有重要学术和社会影响力的学者、知识分子的访谈式介绍。作者始终认为，中国崛起的背后必然是思想的崛起，而这需要有一大批中国思想者的原创性贡献为世界所知。书中所列举的如胡鞍钢、俞可平、何怀宏等学者的人生经历和思想观点，都被列入布鲁金斯学会选编的《当代中国思想家》丛书，进入美国大学的主流课堂。该书不少内容是作者过去几年接受媒体和专栏约稿集结而成的文章编辑而成，反映了作者关于中国智库发展的系统性思考。

作为常年穿梭于太平洋两岸的智库学者，李成始终是中国智库发展的观察者、建言者与合作者。过去几年当中，笔者在工作与研究中都受惠李老师颇多。他对中美两国关系的乐观与想象力，以及为之身体力行的种种努力，都令人心生敬意。