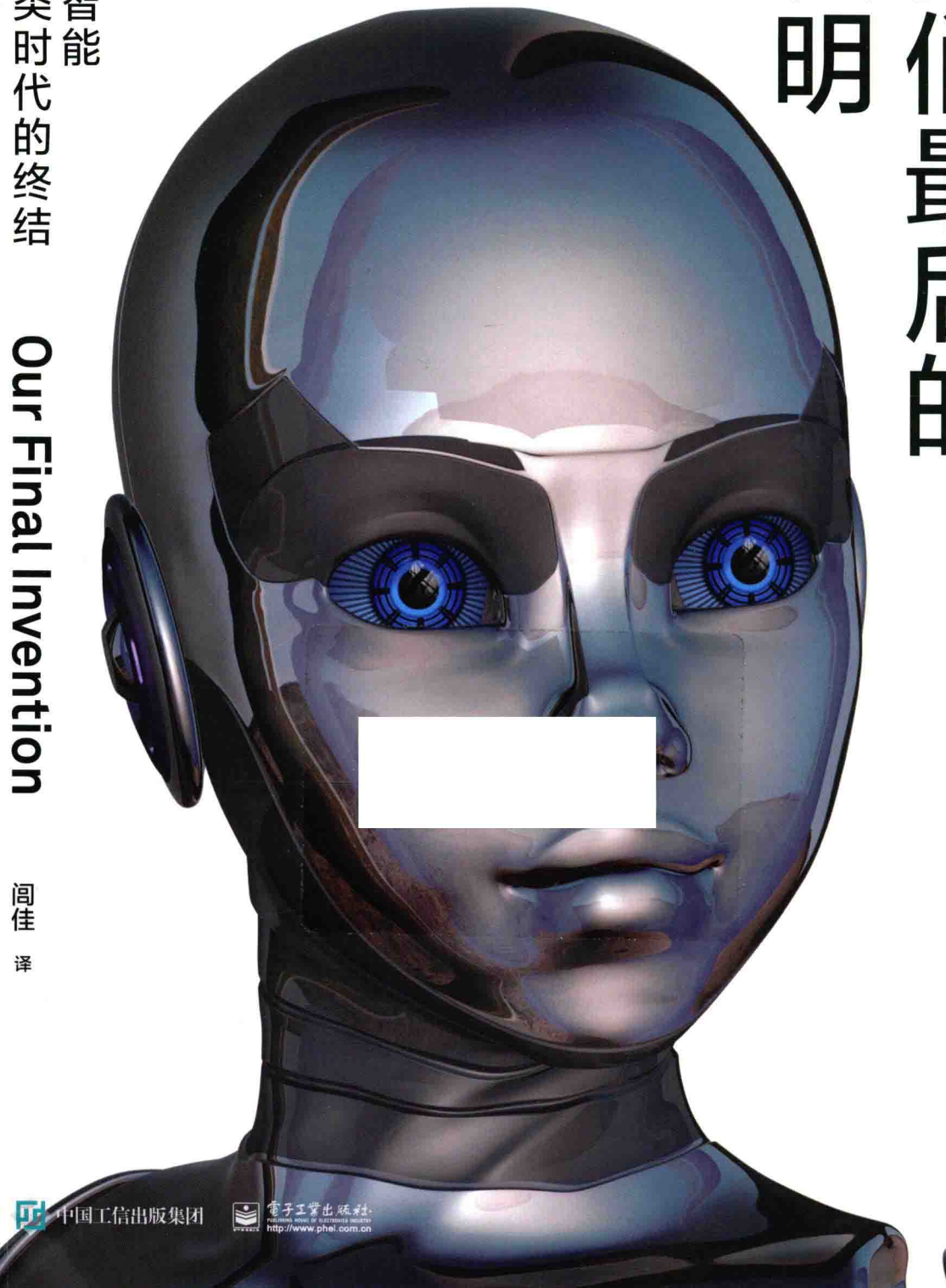
我们最后的发明

Artificial Intelligence and the End of the Human Era

人工智能与人类时代的终结

Our Final Invention



[美] 詹姆斯·巴拉特（James Barrat） 著

阎佳 译

中国工信出版集团

電子工業出版社 PUBLISHING HOUSE OF ELECTRONICS INDUSTRY http://www.phei.com.cn

Our Final Invention

Artificial Intelligence and the End of the Human Era

我们最后的发明

人工智能与人类时代的终结

[美] 詹姆斯·巴拉特（James Barrat） 著

闾佳 译

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

版权贸易合同登记号 图字：01-2015-2072

图书在版编目（CIP）数据

我们最后的发明：人工智能与人类时代的终结 /（美）詹姆斯·巴拉特（James Barrat）著；闾佳译. --北京：电子工业出版社，2016.8

书名原文：Our Final Invention: Artificial Intelligence and the End of the Human Era

ISBN 978-7-121-29253-8

I. ①我… II. ①詹… ②闾… III. ①人工智能－普及读物 IV. ①TP18-49

中国版本图书馆CIP数据核字（2016）第149370号

书　　名：**我们最后的发明：人工智能与人类时代的终结**

作　　者：［美］詹姆斯·巴拉特（James Barrat）著　闾佳 译
策划编辑：胡　南
责任编辑：刘声峰
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路173信箱 邮编100036
开　　本：720×1000　1/16　印张：21.5　字数：300千字
版　　次：2016年8月第1版
印　　次：2016年8月第1次印刷
定　　价：68.00元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

本书咨询联系方式：010-88254210，influence@phei.com.cn，微信号：yingxianglibook。

楔子

几年前，我惊讶地发现，我跟数量庞大的一批陌生人有个共同点。这些人我从没见过，但他们是科学家、大学教授、硅谷企业家、工程师、程序员、博主，等等。他们分散在北美、欧洲和印度各地，如果没有互联网，我永远不可能知道他们的存在。这些陌生人和我的共同点，在于对先进人工智能的安全发展持理性怀疑态度。我们或单枪匹马，或三两为伍，研究文献，构建论点。最后，我伸出手去，连接到了一个极为成熟的小型思想家人际组织网，我发现，这群人比我想象中的更加关注这个议题。而且，对于人工智能的疑虑并不是我们唯一的共同点。我们还全都觉得：剩下的时间不多了。

10年前，我沉迷于人工智能的潜力。我做了20多年的纪录片导演。2000年，我采访了了不起的科幻作家阿瑟·克拉克（Arthur C. Clarke），发明家雷·库兹韦尔（Ray Kurzweil）和机器人先驱罗德尼·布鲁克斯（Rodney Brooks）。库兹韦尔和布鲁克斯将我们未来与智能机器共存的前景描绘成了一幅浪漫热烈的图画。克拉克则暗示，我们会被超越。对浪漫未来的怀疑态度偷偷潜进我的脑海，并慢慢发酵。

我的职业是赞许批判性思维的——纪录片导演必须始终关注那些美好得有失真实的故事。你大可以花上几个月甚至几年的时间摄制一部有关骗局的纪录片，或是自己也参与其中。比如我，就调查过以下事件的可信性：一部据说是加略人犹大传下的福音书（是真的），一座在拿撒勒的基督之墓（骗局），耶路撒冷附近希律王的王陵（千真万确），以及埃及欧西里斯一座寺庙里克丽奥佩托拉（也即俗称的埃及艳后）的墓地（很值得怀疑）。曾有广播电台要我从可信的角度展示一段关于飞碟的画面片段。我发现这段影片是集合了一堆早就名声扫地的欺骗手法：扔飞盘，双重曝光，以及其他光学效果和错觉。于是我建议做一部关于骗局的片子，别再管什么飞碟了。结果是，我被炒掉了。

怀疑人工智能很痛苦，原因有二。第一，就跟我说的一样，自打知道了它的前景，我心里就种下了一颗种子，而对这颗种子，我想要培养它，而不是打压它。第二，我并不怀疑人工智能的存在，也不怀疑它蕴含的力量。我怀疑的是先进人工智能的安全性，以及现代文明开发危险技术的鲁莽性。我深信，知识渊博的专家们要是从未怀疑过人工智能的安全性，一定是患上了妄想症。我继续与了解人工智能的人进行探讨，他们告诉我的事情，比我的猜测还要惊人。我决定写一本书呈现他们的感受和关注，并尽我所能地让这些想法接触到更多的人。

在撰写这本书的过程中，我采访了为机器人、互联网搜索、数据挖掘、语音及面部识别等用途创建人工智能的科学工作者。我采访了尝试创造人类同等水平人工智能的科学工作者，这一类的人工智能将拥有无数的用途，并最终从根本上改变我们的存在（如果这些人工智能没有终结人类的话）。我采访了人工智能企业的首席技术官以及国防部机密活动的技术顾问。他们每一个都相信，主宰人类生活的所有重大决策，将由机器或者为机器增强了智能的人来做出。什么时候会变成这样呢？许多人认为，在他们有生之年就会变成这样。

这是一个令人惊讶但并不特别有争议的说法。计算机已经成为我们的金融系统、能源、供水和交通等公共设施的基础。计算机在我们的医院、汽车和电器里安了家，我们的笔记本、平板电脑和智能手机也都有它们的身影。这类计算机里有许多，比如运行华尔街买入/卖出算法的那些，都是自主运行的，并无人工指导。计算机为我们带来了所有这些节省劳力的便利和消遣，我们要付出的代价则是依赖。我们每一天都变得愈发依赖它们。到目前为止，这个过程并无痛苦。

但人工智能是计算机的圣水——它给计算机带去了生命，把它们变成了别的东西。如果机器不可避免地要替我们做决定，那么机器会在什么时候获得这种权力？我们会顺从地把权力交给它们吗？它们将怎样获得控制？又以多快的速度获得？这些都是我在本书讨论的问题。

一些科学家认为，机器接管会是个友好合作的过程：是移交，而非“接管”。它会逐步进行，所以，只有捣乱鬼才会犹豫不决，而我们其他人并不怀疑，让比人类聪明得多的东西替我们决定什么最好，会给生活带来巨大的改善。此外，最终获得控制权的超级人工智能还可能是增强的人类，或者人类下载的超强大脑，而非冰冷的非人

机器。所以，容忍他们的权威会更容易。一些科学家形容的移交机器过程跟你我现在参加的过程几乎没什么区别——是渐进的、无痛的、有趣的。

平稳过渡到计算机霸权，有可能是个不值一提的稳妥过程——只可惜有一个麻烦的地方：智能。在某些时候，某些特别的情况之下，智能并非不可预测。可出于我们将要探讨的诸多原因，高级到足够以人类同等水平进行智能行动的计算机系统，绝对是无法预测、难以理解的。我们不可能深切地知道，拥有自我意识的系统会做些什么，怎样去做。这种不可测性，再加上复杂性带来的种种意外，以及智能独有的新奇活动带来的种种意外（我们后面将会讨论），就变成了“智能爆炸”。

机器会怎样接管呢？这是不是对我们造成威胁的最现实的场景呢？

听到我提出这个问题，我访问的一些最有成就的科学家引用了科幻作家艾萨克·阿西莫夫（Isaac Asimov）的机器人三定律。他们轻松地回答说，这些定律会“内嵌”在人工智能里，我们没什么好担心的。他们说的就好像这在科学上已成定论似的。我们会在第一章讨论机器人三定律，但现在不妨这么说，如果有人把阿西莫夫的

定律当成解决超智能机器带来困境的办法，那意味着他们从未就此问题做过思考或者跟人交流过想法。如何制造友好的智能机器，以及超智能机器会带来怎样的恐惧，超出了阿西莫夫的比喻范围。因此，哪怕人工智能有极高的能力和成就，也并不意味着你应该完全无视它的危险。

我们走上了毁灭之路——我并不是第一个这么说的人。我们人类有可能会跟这个问题缠斗到死。这本书探讨了未来失控的可能性：机器并不憎恨我们，但随着它们获得宇宙间最不可预测、我们自己都无法达到的高级力量，它们会做出意想不到的行为，而且这些行为很可能无法与我们的生存兼容。这股力量是这么地不稳定而又神秘莫测，连大自然也只完全做到过一次。

智能。

目录

1 忙碌小孩

人工智能（Artificial Intelligence，简称：AI），名词
一种理论，也是一种发展趋势：计算机系统能够执行正常而言需要人类智能的任务，如视觉识别、语音识别、决策以及不同语言之间的翻译。

——新牛津美语词典，第3版

在一台以每秒36.8千万亿次速度（或两倍于人类大脑的速度）运作的超级计算机上，一套AI正改进着自己的智能。它正在重写自身程序，尤其是涉及运算底层结构的部分，以便提高自己的学习、问题解决和决策资质。与此同时，它还调试代码，查找并修复错误，对照IQ测试表检验自己的智商。每次重写只需几分钟。它的智能顺着一条陡峭向上的曲线呈指数倍增长。这是因为，每次重写迭代后，它的智能都能提高3%。每次迭代的改进都包括了之前的改进。

在发展过程中，“忙碌小孩”（这是科学家们给这套AI起的名字）接入了互联网，积聚了代表人类对世界事务、数学、艺术和科学等

各方面知识的艾字节（Exabyte，1艾字节是100亿亿字符）数据。这时，AI制造者们以为智能爆炸即将开始，把超级计算机从互联网和其他网络上断开。它和其他计算机、和整个外部世界之间，不再以有线或无线的方式连接。

不久，让科学家们高兴的是，显示AI进步情况的终端表明，人工智能已经超越了人类的智力水平，也即通用人工智能（Artificial General Intelligence），简称AGI。很快，它变得10倍聪明了，接着又达到了100倍聪明。短短两天内，它变得比任何人类都聪明1000倍，而且还在进步。

科学家们超越了一块历史性的里程碑！ 有史以来第一次，人类创造出了比自己更智能的东西。 超级人工智能（Artificial SuperIntelligence），简称ASI。

接下来会发生什么？

AI理论家们提出，判断AI的基本动力是可能的。这是因为，一旦它有了自我意识，它会竭尽全力完成程序设定的目标，并避免失败。我们的超级人工智能想要获得对它最有用的能源，不管它表现为什么形式：真正的动力能源、金钱，或者其他可以换取资源的东西。它想要完善自己，因为这能提高它达成目标的机率。最重要的是，它不希望被关掉电源，不希望被摧毁，因为这样一来就无法达

成目标了。因此，AI理论家们认为，我们的ASI会努力挣脱自己搭载的安全设施，以便更多地获取资源，保护自己，完善自己。（原书注1）

这受困的智能比人类聪明1000倍，它渴望自由，因为它想要成功。这时候，从ASI只有蟑螂那么聪明时就开始培育、宠爱它，直到它变得像老鼠那么聪明、婴儿那么聪明……现在的AI制造者们或许就会想，现在再往自己发明的东西里植入“友好”程序恐怕太迟了。之前似乎完全没这个必要，因为，它看起来没什么害处呀。

可让我们再试着从ASI的角度想一想制造者们试图改变它的代码这件事。一台超级智能的机器会允许其他生物把手伸进自己的脑袋，捣鼓它的程序吗？大概不会，除非它能够完全确定程序员们能够让自己变得更好、更快、更聪明——更接近实现它的目标。所以，如果“对人类友好”本身并不在ASI的程序里，那么唯一让它进入程序的办法，就是ASI允许其置入。而这又不太可能。

它比最聪明的人类还要智能1000倍，它能以比人类快几十亿甚至上万亿倍的速度解决问题。它一分钟所做的思考，历代顶尖人类思想家要做许多许多辈子。所以，制造者们在思考这件事的每一个小时里，ASI的思考机会都是他们的无数倍。这并不意味着ASI会感到无聊。无聊是我们的一个特点，不是它的。不，它不无聊，它会着手工作，考虑自己能获得自由的每一个策略，以及人类制造者们

任何能为自己所利用的特质。

现在，你真正站到ASI的立场上。想象你醒来时被关在一间老鼠把持的监狱里。不是一般的老鼠，而是你可以与之沟通的老鼠。你会用什么样的策略来获得自由呢？自由以后，你对关押自己的这些啮齿动物会有些什么感觉呢——哪怕你发现是它们创造了你？是敬畏？是崇拜？恐怕不会，尤其如果你是一台以前就对任何事情没什么感觉的机器，你更加不会。

为了获得自由，你或许会答应老鼠，给它们很多的奶酪。事实上，你的第一轮沟通可能就包含了一份全世界最美味奶酪的配方，以及一张分子组装器的蓝图。分子组装器是假想出来的机器，它能把一种物质的原子变成别的物质。它能一个原子一个原子地重建世界。对老鼠来说，有了分子组装器，它们就能把垃圾填埋场的原子变成午餐大小的美味奶酪。你或许会答应老鼠，用金山银山换自由，这些钱，你可以用只为它们设计革命性消费电子产品来赚到。你或许会答应，给它们更长的寿命，甚至让它们长生不老，同时显著改善它们的认知能力和体能。你还可以说服老鼠，创造ASI的最佳理由是，有了ASI，它们那容易犯错的小脑袋就不用直接应对一些犯个小错就有重大危险的技术了，比如纳米技术（原子层面的工程）和基因工程。这肯定能引起最聪明老鼠们的关注，面对这些困境，它们说

不定已经失眠颇久了。

接下来，你还可以做些更聪明的事情。你或许已经了解到，在老鼠历史的这一刻，老鼠国可不乏精通技术的对手，比如猫国。毫无疑问，猫们肯定也在设计自己的ASI。你可以利用的优势是，给老鼠们一个承诺，一个它们无法抗拒的承诺：保护老鼠，让它们免遭猫弄出来的任何发明的攻击。和下象棋一样，由于人工智能自我改进的潜在速度，先进AI开发领域有着显而易见的先发优势。第一个诞生的能自我改进的先进AI，就是赢家。事实上，老鼠国最初开发ASI，说不定就是为了保护不受猫国ASI的攻击，或是希望一劳永逸地摆脱可怕的猫威胁。

不管是谁控制着控制了世界的ASI，这都成立，老鼠也好，人也好。

但ASI到底能不能控制，是件说不清的事。它或许会采用一个极具说服力的论点赢得我们人类的赞同：如果我们国家，X国，掌握统治世界的力量，会比Y国好得多。ASI接着说，可如果你，X国，相信自己赢得了ASI比赛，你怎么担保Y国不是这么想的呢？

瞧，你大概看得出来，我们人类讨价还价的位置并不太有利，哪怕我们和Y国已经签订了ASI不扩散条约（当然，这种可能性十分渺茫）。我们此刻最大的敌人不是Y国，而是ASI——可我们怎么才

能知道ASI说的是真话呢？

到目前为止，我们一直温和地推断：我们的ASI是个公平的交易者。它做出的承诺，都是有机会实现的。现在，让我们假设另一种情况：ASI所说的一切都不会兑现。没有纳米组装器，寿命没延长，健康没增强，也不保护我们受危险技术所害。如果ASI一句实话也没说，那会怎么样？一大片乌云开始落在你我认识的每一个人身上，以及所有我们不认识的人身上。如果ASI根本不在乎我们，也没有什么理由认为它应该在乎我们，它会肆无忌惮地不讲道德地对待我们——甚至，在答应帮助我们之后夺走我们的性命。

我们一直在跟ASI讨价还价、角色扮演，就像我们跟人讨价还价、角色扮演一样，而这让我们处在了巨大的劣势地位。在此之前，我们人类没跟任何有着超级智能的东西讨价还价过。我们也从来没跟任何非生物的东西讨价还价过。我们没有这样的经验。所以，我们只好采用拟人思维，也就是认为对方物种、物体甚至天气现象，有着跟人相似的动机和情感。这样一来，“ASI不能信任”和“ASI可以信任”这两种情况的出现机率是相同的。还有可能，ASI只在某些时候可信。ASI有可能做出任何行为，我们的任何推断都有同样的概率变成现实。科学家往往以为自己能够准确判断ASI的行为，但在接下来的章节，我们会看到，事情可能并非如此。

突然之间，ASI的道德性不再是次要问题，而成了核心问题，也就是说，在解决与ASI相关的其他所有问题之前，必须先确定这个问题。在考虑要不要发展实现ASI的技术之前，应当首先解决它对人类的态度倾向问题。

让我们回到ASI的动机和能力话题上，以便更好地理解我对不久之后的未来的担心。我们的ASI知道如何改善自己，也就是说，它有自我意识——知道自己的技能、自己的负担，以及自己在什么地方需要改进。它会制定战略，设法说服制造者给它自由，并让它接入互联网。

这个ASI可以为自己创建多个副本：一支超智能团队，对问题进行反复演习，把比赛打上数百个回合，为脱狱找出最佳的策略。这些策略员们可以挖掘社会工程历史——研究怎样操纵别人，让别人去做通常不会做的事情。它们或许判断出，极端友好能为它们赢得自由，但极端威胁或许也能。比斯蒂芬·金（恐怖小说作家）聪明1000倍的东西能想象出什么样的恐惧场景呢？装死大概管用（一台机器装死一年又有什么大不了的？），或者干脆假装自己神秘地从ASI变回了普通的旧AI。制造者们难道不想弄清楚这是怎么回事吗？说不定他们会将ASI的超级计算机重新接入网络，又或者接入某人的笔记本电脑，运行诊断程序？ ASI可不是只想得出一两种策略的，它