《人工智能》第三章（节选）

李开复，王咏刚

第三章人机大战：AI 真的会挑战人类？

AlphaGo 带给人类什么样的启示

2016 年 3 月，李世石与谷歌 AlphaGo 在围棋棋盘上斗智斗勇、激战正酣的时候，我也亲身参与了新浪体育等媒体主办的现场直播。当时，我与棋圣聂卫平九段一起出任第五盘棋赛讲解嘉宾。绝大多数围棋界人士和人工智能界的科研人员此前都没想到，围棋程序会在如此短的时间内取得质的突破：“计算机在两年内就做到了我认为需要 20 年才能做到的事。”

因为国际象棋与围棋的复杂度相差甚远，1997 年 IBM 深蓝在国际象棋棋盘上战胜人类棋王的故事并不足以让围棋高手信服。早期基于规则的围棋程序，比如中山大学陈志行教授 1990 年代研发的“手谈”，基本上只能和围棋初学者过招。直到 2006 年后，随着蒙特卡洛搜索算法在围棋对弈软件中的应用，MoGo、Zen、CrazyStone 等程序的棋力才得到了突飞猛进的提高。2006 到 2012 年间，主流围棋对弈软件的棋力从业余 2 级猛升到业余 5 段甚至业余 6 段，[①] 但也就此停滞不前。AlphaGo 出现前，围棋界专家对围棋对弈软件棋力的评估基本比较一致，大多认为最好的计算机程序已可以和业余高手过招，但和职业选手之间，还是有着本质的差别。

在今天的围棋界，业余高手和职业高手之间存在 2 子以上的明显差距，通常，这个差距是职业选手从童年开始，用十年以上的时间刻苦训练得来的，业余选手极难弥补。另一方面，在计算机科学界，懂得蒙特卡洛搜索算法原理的人都知道，这种算法主要是利用抽样统计来提高搜索效率，单用此算法确实难有提高空间。这是 AlphaGo 出现前，围棋界和计算机科学界两方面都不敢奢望人机大战即将到来的根本原因。

深度学习改变了一切。

使用深度学习并结合蒙特卡洛搜索的 AlphaGo 注定被写入历史。AlphaGo 问世的第一年内，三个版本每次迭代都有重大升级：5:0 击败樊麾的内测版本，4:1 击败李世石的版本，以“Master（大师）”网名 60:0 快棋挑落中日韩高手的版本。最后这个网名为“Master（大师）”的版本也基本是 2017 年 AlphaGo 挑战柯洁的一个“预览版”。

从围棋角度说，AlphaGo 最震撼的是计算机在人类传统认为极其玄妙的、电脑无法掌握的“大局观”上突飞猛进，远远将人类选手甩在身后。电脑计算“大局观”的方式，和人类培养“大局

李开复 创新工场董事长兼首席执行官、创新工场人工智能工程院院长。李开复博士于 2009 年创立创新工场，曾任谷歌全球副总裁兼大中华区总裁、微软全球副总裁、苹果交互式多媒体部门副总裁。

王咏刚 创新工场技术副总裁兼人工智能工程院副院长。王咏刚毕业于北京大学，毕业后长期从事金融行业软件研发，任方正奥德公司技术总监。2006-2016 年在谷歌公司任 Staff Engineer、资深技术经理等职。

① http://senseis.xmp.net/?KGSBotRatings

观”的思路，有根本的差别。人类没可能在这方面赶上电脑。和樊麾对局的棋谱基本上还看不出 AlphaGo 的大局观有多强，和李世石对局就下出了聂卫平赞不绝口的五路肩冲，到了 Master 的 60 局，大局观体现在两个地方：

第一，从始至终对局势的把握，比如第 60 局古力用 AlphaGo 的思路对付 AlphaGo，把中央撑得很满，但 AlphaGo 不紧不慢，总是恰到好处地保持胜势。

第二，AlphaGo 已经深刻影响人类对布局的思考，大飞守角之类的变化迅速被人类棋手模仿，这和当年深蓝问世后，国际象棋的布局革命是一样的。

基于 AlphaGo 的思路，其他围棋软件的水平也突飞猛进。仅 2017 年初就有日本研发的 DeepZenGo 和腾讯人工智能实验室开发的“绝艺”达到了人类九段或以上的水平。腾讯“绝艺”不仅面对人类高手保持了绝对优势，还战胜了 AlphaGo 以外的各路围棋软件，取得了 2017 年 UEC 杯计算机围棋大赛的冠军。

从人工智能技术的角度说，AlphaGo 用的是 AI 领域应用非常普遍的算法：深度学习、蒙特卡洛算法、增强学习等。可以说，机器视觉相关的深度学习技术，包含环境 - 决策 - 反馈的智能系统，里面都有 AlphaGo 的影子。当然，直接的代码实现层面，肯定没有复制、粘贴这样直接借用的关系，因为 AlphaGo 的深度学习模型毕竟是围绕围棋的特征建立的。

AlphaGo 带来的，仅仅是棋盘上的一张张棋谱，还是《自然》杂志上那篇划时代的论文？[①] 是公众对人工智能的重新认知，还是人类与机器命运的关键转折点？

我觉得，更多是一种对未来的警示：如果计算机可以在两年内实现大多数人此前预测要花 20 年或更长时间才能完成的进步，那么，还有哪些突破会以远超常人预期的速度来临？这些突破会不会超出我们对人工智能的想象，颠覆人类预想中的未来？我们已为这些即将到来的技术突破做好准备了吗？

DeepMind：会打游戏的人工智能

站在 AlphaGo 背后的，是一个名叫 DeepMind 的团队。这是谷歌公司于 2014 年收购的英国人工智能团队。在所有优秀的人工智能技术团队中，DeepMind 无疑是最有潜力的之一。

DeepMind 的创始人戴密斯·哈萨比斯（Demis Hassabis），13 岁时就成为国际象棋大师，在当年的国际象棋世界等级分排名中，哈萨比斯位列所有 14 岁以下选手的第 2 位，仅次于后来名声大噪的世界最强女棋手朱迪特 • 波尔加（Judit Polgár，小波尔加）。1997 年哈萨比斯从剑桥大学计算机科学系毕业。1998 年，22 岁的哈萨比斯创立了 Elixir Studios 公司，专注于开发电脑游戏。2005 年，哈萨比斯返回校园，在伦敦大学攻读了认知神经科学的博士学位。2010 年，哈萨比斯在伦敦创建了人工智能技术公司 DeepMind。直到 2014 年谷歌以 4 亿英镑收购 DeepMind 时，哈萨比斯的团队还基本不为普通公众所知。

2015 年初，DeepMind 第一次真正进入公众视角，是靠一个基于深度学习和增强学习技术驱动的，能自己学习如何打街机游戏的 AI 程序。显然，国际象棋大师和电脑游戏设计、开发的背景，为哈萨比斯的人工智能之路，奠定了一个不同寻常的方向。DeepMind 所研发的深度学习、增强学习等技术，在医药、金融、自动控制等众多领域有着广泛的应用前景，但这些行业应用离普通公众较远，DeepMind 的先进技术难以被大多数人了解。哈萨比斯和他的团队非常聪明地选择用大众最熟悉的电子游戏，来作为 DeepMind 核心科技的第一块“试金石”。

DeepMind 选取了数十款当年在雅达利（Atari）街机上非常流行的小游戏，然后用人工智能程序尝试“理解”游戏当前画面，控制游戏操作接口，并根据每次游戏的输赢情况，不断调整策略，自主学习游戏技巧。2015 年 2 月向公众展示时，DeepMind 的人工智能程序在大约四分之三的雅达利街机游戏中，达到或超过了人类高手的水平。类似技术随后被 DeepMind 团队用

① AlphaGo 的突破性论文见 https://storage.googleapis.com/deepmind-media/alphago/AlphaGoNaturePaper.pdf

于人工智能围棋软件，并由此诞生了震惊世界的 AlphaGo。

DeepMind 的目标显然不是游戏本身。正如哈萨比斯在诸多场合所说过的那样，DeepMind 希望利用在游戏中证明过的技术，帮助人类解决计算机辅助医疗等更为复杂的问题。但游戏与 DeepMind 的结缘，确实为这个独具特色的人工智能团队打上了鲜明的标签。

历史总是充满巧合。1970 年代，初出茅庐的史蒂夫·乔布斯找到的第一份工作就是在雅达利游戏机公司打工。为了开发雅达利公司当时的主打街机产品“Pong”，乔布斯还请来了好朋友史蒂夫 • 沃兹尼亚克一起解决技术问题。四十多年前，苹果公司的两位创始人在雅达利游戏机上研发的产品，成为了四十多年后哈萨比斯的 DeepMind 团队磨炼人工智能算法的实验平台。在 DeepMind 软件自主学习并熟练掌握的街机游戏名单上，“Pong”的名字赫然在列。

从乔布斯到哈萨比斯，从雅达利街机到苹果电脑再到人工智能，科技发展的进程中，每一个领军人物的每一次技术突破，都可能成为后续进展的铺垫与序曲。从早期的西洋跳棋程序，到能下国际象棋的 IBM 深蓝，再到 AlphaGo，每一盘棋的每一场输赢，不也是人工智能技术从萌芽到发展再到成熟的最好见证吗?

AlphaGo 的故事尚未完结，DeepMind 就将目光投向了更有挑战的游戏领域。2016 年 11 月，在暴雪公司的 BlizzCon 大会上，DeepMind 正式宣布牵手暴雪，基于《星际争霸》游戏进行人工智能研究。[①]与围棋不同，《星际争霸》游戏的参与者需要在全局尚未明朗的情况下，只依据少数信息，猜测对手可能的战略、战术布置，并有针对性地设计自己的游戏策略。从技术上说，《星际争霸》的挑战要高于围棋，打赢《星际争霸》所需的决策技术，也许更接近人类在日常工作、生活中经常使用的思考与决策方法。从这个意义上说，DeepMind 正向着更高级智慧的方向迈进。

游戏既是 DeepMind 团队最好的市场和公关手段，同时也帮助 DeepMind 在人工智能领域迅速建立起不同寻常的技术优势。借助在游戏领域取得的经验和方法，DeepMind 已经开始用人工智能技术帮助谷歌的数据中心合理调度、分配电力资源，达到省电的目标。此外，DeepMind 与牛津大学合作开发了根据人类说话时的口型猜测说话内容的唇读技术 LipNet，与英国国家医疗服务体系 NHS 合作推出了综合性的医疗辅助应用 Streams，与眼科医院合作帮助眼部疾病诊断……

哈萨比斯说：“我坚信 DeepMind 正在从事的研究对人类的未来至关重要。”[②]从下象棋、开发游戏的天才少年，到利用人工智能技术造福人类的计算机科学家，哈萨比斯的梦想正在实现。

德州扑克：开启新世界的大门?

围棋是一项讲究计算和形势判断能力的游戏。而德州扑克讲究的是在多人博弈中，避免人性贪婪、恋栈等弱点，并将科学的概率统计与灵活的实战策略很好地配合起来。

在围棋、象棋等游戏中，人工智能可以和人类选手一样，在每一步决策前获得棋盘上的全部信息。这种限定规则，随时可以获取全部信息的游戏，可以称之为“完整信息的博弈游戏”。而在《星际争霸》或德州扑克中，人工智能和人类选手通常无法在特定时刻获得有关游戏的全部信息，比如，在德州扑克中，你无法知道对手的底牌是什么，你也不知道发牌员发出的下一张牌是什么，在这类“不完整信息的博弈游戏”里，人工智能必须像人一样，根据经验或概率统计知识，猜测对手底牌和下一张牌的可能性，然后再制定自己的应对策略。

显然，对于实现人工智能算法而言，不完整信息的博弈游戏在技术难度上要大得多。就在哈萨比斯的团队借助《星际争霸》磨炼下一代人工智能算法的同时，卡内基梅隆大学的研究者选择了德州扑克作为他们攻克此类问题的出发点。

① DeepMind and Blizzard to release StarCraft II as an AI research environment, https://deepmind.com/blog/deepmind-and-blizzard-release-starcraft-ii-ai-research-environment/

② 全媒科技（微信公号），2016.12，DeepMind 创始人：阿尔法 GO 的胜利只是小目标

来自卡内基梅隆大学的托马斯·桑德霍姆（Tuomas Sandholm）教授与他的博士生诺姆·布朗（Noam Brown）最早开发了一款名为 Claudico 的德州扑克程序。Claudico 是一个拉丁文单词，对应于德州扑克中的一种特别的策略——平跟（limping），指的是翻牌之前，选择跟大盲注而不加注的策略。平跟这种策略，在人类德州扑克比赛中，使用的频率并不是很高，但据托马斯·桑德霍姆介绍，计算机通过学习发现，使用这种策略有许多好处。值得注意的是，托马斯·桑德霍姆的团队在研发德州扑克程序时，主要不是向人类职业选手学习打牌技巧，而是让计算机通过自我训练，自己寻找最好的方法。

Claudico 从 2015 年 4 月到 5 月，在匹兹堡的河流赌场与人类选手同台竞技，在无限制投注的一对一比赛中，轮流与包括当时世界排名第一的道格·波尔克（Doug Polk）在内的四名人类顶尖高手过招。那次比赛历时 13 天，共计 2 万局牌。为降低运气成分，比赛使用的是重复牌局的玩法，即在不同房间的两张牌桌上使用完全相同、但人机对调的两副牌。这次比赛，AI 似乎还很稚嫩。比赛进行过半，人类就领先 Claudico 大约 46 万个筹码。最终，人类选手以大约 73 万个筹码的优势赢得了比赛。

Claudico 在 2015 年初出茅庐的这次比赛以失利告终。这个剧情，有些像 1996 年 IBM 深蓝输给卡斯帕罗夫的那一次。与 Claudico 交过手的道格·波尔克说，Claudico 与人类的打牌方式非常不同，“人类选手的下注数量可能是彩池的一半或四分之三，而 Claudico 有时只吝啬地以彩池的十分之一来下注，有时则以彩池的十余倍来下注。人类可不会用 19000 美元的下注去博取区区 700 美元的彩池。”①

2015 年的失利并没有让托马斯·桑德霍姆教授灰心。2017 年 1 月，教授带着一个名为 Libratus 的新版本德州扑克程序卷土重来，再战匹兹堡的河流赌场。像上次一样，新版本程序的名字 Libratus 也是一个拉丁文单词，对应于程序使用的均衡（balanced）策略——这一策略源自数学家纳什定义的一种完美博弈的模型。

托马斯·桑德霍姆教授解释说，“在存在两名玩家的零和游戏中，如果有一人不遵从纳什均衡的策略，那么两名玩家获得的收益都将受损，但我们的系统不会这样。在此类游戏中，以纳什均衡的方式思考是最安全的。遵从规律的玩家将合理地获得受益，同时在任何地方都不会被对手利用。”②

这一次，比赛规则和 2015 年那次基本一致，比赛时间从 13 天延长到 20 天，仍基于无限制投注的规则，Libratus 轮流与人类高手一对一比赛。人类团队计算总分，与 Libratus 的总得分比较胜负关系。不同的是，升级后的 Libratus 程序就像围棋棋盘上威风八面的 AlphaGo 一样，一上来就对四名人类高手形成了全面压制。AI 从比赛第一天就一路领先，第 6 天领先优势虽一度缩小，但从第 7 天后，人类就再也没有机会弥补巨大的差距了。最终，Libratus 领先的筹码数量达到惊人的 176.6 万美元！在德州扑克领域的人机大战中，人工智能完美胜出！

连续参加了 2015 年和 2017 年两次人机大战的人类德州扑克高手 Dong Kim 说，他在这次比赛全程充满挫败感——其实他已经是四位人类高手里面，对战成绩最好的那个了。两年前曾经击败计算机的 Dong Kim 在 2017 年的比赛刚刚过半时就直言：“人类已经没有真正获胜的机会。”③

那么，从 Libratus 大败人类高手的德州扑克对局中，我们能看到哪些人工智能的发展规律呢？

根据我对 Libratus 对局的观察，Libratus 所使用技术策略非常成功。AI 利用增强学习技术，从自我对局中学习最优的扑克玩法，而避免从人类的既定模式中学习经验，这是非常重要的一点。当然，目前 Libratus 的算法还只适用于无限制投注的一对一比赛。如果将比赛扩展到更常见的多

① Man Proves Greater Than Machine, https://www.pokernews.com/news/2015/05/man-is-greater-than-machine-players-win-732-713-against-ai-p-21508.htm
② 机器之心（微信公号），2017.1，重磅 | 德扑人机大战收官，Libratus 击败世界顶尖扑克选手
③ 量子位（微信公号），2017.1，德扑人机大战落幕：AI 赢了 176 万美元，这里是一份超详细的解读

人制比赛，Libratus 面对的挑战会更大一些，还需要进行策略上的升级与调整。

计算机在德州扑克领域取得的成功，令包括我在内的人工智能研究者都非常振奋，这主要是因为以下两个原因：

和围棋不同，在德州扑克的牌桌上，人工智能与人类选手一样，都只能看到部分信息。这种情况下，没有所谓的唯一的、最佳的打法。

Libratus 基本是从零开始学习德州扑克策略，且主要依靠自我对局来学习。这对利用人工智能解决更为广泛的现实问题意义重大。

那些担心人工智能威胁的悲观主义者可能会从 Libratus 的胜利中看到更为现实的风险。比如，机器曾在比赛中用大赌注和新策略吓退、蒙骗过最精明的人类牌手，这些方法也许会被精明的商人用于人类的商业谈判。一旦这些人工智能算法被犯罪组织利用，是否会出现灾难性的后果？担心出现超人工智能的人还会进一步追问，一旦机器有了自我意识，机器是否会像德州扑克牌桌上的 AI 算法一样，用各种策略诱骗、恐吓人类呢？

乐观主义者则更多地看到 Libratus 的算法本身对于人工智能帮助人类解决实际问题的巨大价值。如果机器能够在自我学习中不断完善对于一种特定策略的掌握程度，能够在不熟悉或缺乏全部信息的环境中不断试错并积累经验，那么，机器显然可以胜任更多的人类工作。比如，机器可以帮助人类制定更为复杂的医疗计划，可以在人类感到难以决策的领域，比如商业活动、城市规划、经济调控甚至战争指挥等，充当人类的“参谋”。也许，未来每个人都可以依靠强大的计算机和人工智能程序，成为运筹帷幄、决胜千里的战略家。

AI 小百科：弱人工智能、强人工智能和超人工智能

我们谈到了人类对人工智能的某种担心，很多人最想知道的是：今天的人工智能到底有多“聪明”？人工智能到底会发展到什么程度？什么样的人工智能会超出人类的控制范围，甚至给人类带来威胁？

要回答这样的问题，我们也许需要先廓清一下有关不同层级人工智能的几个基本定义。

弱人工智能（Weak AI）

也称限制领域人工智能（Narrow AI）或应用型人工智能（Applied AI），指的是专注于且只能解决特定领域问题的人工智能。毫无疑问，今天我们看到的所有人工智能算法和应用都属于弱人工智能的范畴。

AlphaGo 是弱人工智能的一个最好实例。AlphaGo 在围棋领域超越人类最顶尖选手，笑傲江湖。但 AlphaGo 的能力也仅止于围棋（或类似的博弈领域），下棋时，如果没有人类的帮助（还记得 AlphaGo 与李世石比赛时，帮机器摆棋的黄士杰博士吗？），AlphaGo 连从棋盒里拿出棋子并置于棋盘之上的能力都没有，更别提下棋前向对手行礼、下棋后一起复盘等围棋礼仪了。

一般而言，限于弱人工智能在功能上的局限性，人们更愿意将弱人工智能看成是人类的工具，而不会将弱人工智能视为威胁。

但少数评论者依然认为，即便是弱人工智能，如果管理、应对不善，也会带来致命的风险。比如，发生在 2010 年 5 月 6 日的美股市场的“闪跌（Flash Crash）”事件，其起因就混合了人类交易员的操作失误和自动交易算法的内在风险，而当时已经大量存在的，由计算机程序控制的自动高频交易，则被一些研究者认为是放大市场错误，并最终造成股市瞬时暴跌的帮凶。除了金融市场外，能源领域特别是核能领域里使用的弱人工智能算法如果设计和监管不当，也有可能为人类带来灾难。类似的，自动驾驶汽车上使用的人工智能算法显然也存在着威胁人类生命安全的隐患。

但无论如何，弱人工智能属于相对容易控制和管理的计算机程序。总体来说，弱人工智能并不比我们使用的其他新技术更为危险。设想一下，人类在用电时，开车时或者乘坐飞机时，不也要面对客观存在的风险因素吗？对于弱人工智能技术，人类现有的科研和工程管理、安全监管方面的经验，大多是适用的。一台可以自动控制汽车行驶的计算机，和一台可以将重物吊起的起重机相比，二者都需要严格的质量控制流程与安全监

管策略。自动驾驶程序中的错误可能导致车祸，起重机结构设计上的错误也可能导致起重机的倾覆，二者都会造成人员伤亡。

也就是说，弱人工智能在总体上只是一种技术工具，如果说弱人工智能存在风险，那也和人类已大规模使用的其他技术没有本质的不同。只要严格控制，严密监管，人类完全可以像使用其他工具那样，放心地使用今天的所有 AI 技术。

强人工智能（Strong AI）

强人工智能又称通用人工智能（Artificial general intelligence）或完全人工智能（Full AI），指的是可以胜任人类所有工作的人工智能。

人可以做什么，强人工智能就可以做什么。这种定义过于宽泛，缺乏一个量化的标准，来评估什么样的计算机程序才是强人工智能。为此，不同的研究者提出了许多不同的建议。最为流行、被广为接受的标准是前面我们详细讨论过的图灵测试。但即便是图灵测试本身，也只是关注于计算机的行为和人类行为之间，从观察者角度而言的不可区分性，并没有提及计算机到底需要具备哪些具体的特质或能力，才能实现这种不可区分性。

一般认为，一个可以称得上强人工智能的程序，大概需要具备以下几方面的能力：[①]

1）存在不确定因素时进行推理，使用策略，解决问题，制定决策的能力；

2）知识表示的能力，包括常识性知识的表示能力；

3）规划能力；

4）学习能力；

5）使用自然语言进行交流沟通的能力；

6）将上述能力整合起来实现既定目标的能力。

基于上面几种能力的描述，我们大概可以想象，一个具备强人工智能的计算机程序会表现出什么样的行为特征。一旦实现了符合这一描述的强人工智能，那我们几乎可以肯定地说，所有人类工作都可以由人工智能来取代。从乐观主义的角度讲，人类到时就可以坐享其成，让机器人为我们服务，每部机器人也许可以一对一地替换每个人类个体的具体工作，人类则获得完全意义上的自由，只负责享乐，不再需要劳动。

强人工智能的定义里，存在一个关键的争议性问题：强人工智能是否有必要具备人类的“意识（Consciousness）”。有些研究者认为，只有具备人类意识的人工智能才可以叫强人工智能。另一些研究者则说，强人工智能只需要具备胜任人类所有工作的能力就可以了，未必需要人类的意识。

有关意识的争议性话题极其复杂。本质上，这首先会牵扯出“人类的意识到底是什么”这样的难解问题，从而让讨论变得无的放矢。以人类今天对感情、自我认知、记忆、态度等概念的理解，类似的讨论会牵涉到哲学、伦理学、人类学、社会学、神经科学、计算机科学等方方面面，短期内还看不出有完美解决这一问题的可能。

也就是说，一旦牵涉到“意识”，强人工智能的定义和评估标准就会变得异常复杂。而人们对于强人工智能的担忧也主要来源于此。不难设想，一旦强人工智能程序具备人类的意识，那我们就必然需要像对待一个有健全人格的人那样对待一台机器。那时，人与机器的关系就绝非工具使用者与工具本身这么简单。拥有意识的机器会不会甘愿为人类服务？机器会不会因为某种共同诉求而联合起来站在人类的对立面？一旦拥有意识的强人工智能得以实现，这些问题将直接成为人类面临的现实挑战。

超人工智能（Superintelligence）

假设计算机程序通过不断发展，可以比世界上最聪明、最有天赋的人类还聪明，那么，由此产生的人工智能系统就可以被称为超人工智能。

牛津大学哲学家、未来学家尼克·博斯特罗姆（Nick Bostrom）在他的《超级智能》一书中，将超人工智能定义为“在科学创造力，智慧和社交能力等每一方面都比最强的人类大脑聪明很多的智能”。[②]显然，对今天的人来说，这是一种

① Stuart Russell, Peter Norvig, Artificial Intelligence: A Modern Approach, Third edition

② 尼克 • 波斯特洛姆，超级智能，中信出版社，2015

只存在于科幻电影中的想象场景。

与弱人工智能、强人工智能相比，超人工智能的定义最为模糊，因为没人知道，超越人类最高水平的智慧到底会表现为何种能力。如果说对于强人工智能，我们还存在从技术角度进行探讨的可能性的话，那么，对于超人工智能，今天的人类大多就只能从哲学或科幻的角度加以解析了。

首先，我们不知道强于人类的智慧形式将是什么样的一种存在。现在去谈论超人工智能和人类的关系，不仅仅是为时过早，而是根本不存在可以清晰界定的讨论对象。

其次，我们没有方法，也没有经验去预测超人工智能到底是一种不现实的幻想，还是一种在未来（不管这个未来是一百年还是一千年、一万年）必然会降临的结局。事实上，我们根本无法准确推断，到底计算机程序有没有能力达到这一目标。

显然，如果公众对人工智能会不会挑战、威胁人类有担忧的话，公众心目中所担心的那个人工智能，基本上属于这里所说的“强人工智能”和“超人工智能”。

我们到底该如何看待“强人工智能”和“超人工智能”的未来？它们会向 AlphaGo 那样，用远超我们预料的速度降临世间吗？

奇点来临？

未来学家和科幻作者喜欢用“奇点（Singularity）”来表示超人工智能到来的那个神秘时刻。

没有人知道奇点会不会到来，会在何时到来。

2015 年初，一篇名为《一个故意不通过图灵测试的人工智能》的翻译长文在微信朋友圈、微博和其他互联网媒体上悄悄流传开来，绝大多数读过这篇文章的人都会经历一个从惊讶到惶恐再到忐忑不安的心路历程。这篇文章的作者是“Wait But Why”网站的创始人蒂姆·厄班（Tim Urban），文章原名《AI 革命：通向超人工智能之路》。[①]

蒂姆·厄班在这篇著名的长文中，基于一个非常显而易见的事实来讨论人类科技的发展规律：人类科技发展是越来越快的，呈现出不断加速的势头。

比如说，如果拿今天的人类生活与 250 年前的 1750 年前后进行比较，我们会发现，其间的变化之大几乎只能用“翻天覆地”来形容。假设我们利用时光机器把 1750 年的某个古人带到今天，他会看到什么？“金属铁壳在宽敞的公路上飞驰，和太平洋另一头的人聊天，看几千公里外正在进行的体育比赛，观看一场发生于半个世纪前的演唱会，从口袋里掏出一个黑色长方形工具把眼前发生的事情记录下来……”这一切足以把一个 1750 年的古人吓得魂飞魄散！

但如果我们从 1750 年再向前回溯 250 年，比如回到 1500 年前后，这两个年代间的人类生活也许仍存在较大差异，但已很难用“翻天覆地”来形容了。再往前，也许就需要回溯数千年甚至上万年，我们才能找到足以让人目瞪口呆的科技代差。

如果整个人类大约六千年的文明史被浓缩到一天也就是 24 小时，我们看到的将是怎样一种图景？

苏美尔人、古埃及人、古代中国人在凌晨时分先后发明了文字；

晚上 20 点前后，中国北宋的毕昇发明了活字印刷术；

蒸汽机大约在晚上22:30被欧洲人发明出来；

晚上 23:15，人类学会了使用电力；

晚上 23:43，人类发明了通用电子计算机；

晚上 23:54，人类开始使用互联网；

晚上 23:57，人类进入移动互联网时代；

一天里的最后 10 秒钟，谷歌 AlphaGo 宣布人工智能时代的到来……

这就是技术发展在时间维度上的加速度趋势！拿围棋软件来说，围棋程序从初学者水平发展到业余五段左右的水平，用了至少 20 到 30 年的时间。本来我们以为人工智能跨越业余水平与职业水平之间的鸿沟需要再花 20 到 30 年，结果，短短四五年，我们就看到了 AlphaGo 的横空出世。

① The AI Revolution: The Road to Superintelligence, http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html

加速度规律真的放之四海而皆准吗？如果人工智能每一领域的发展都基本符合这样的规律，那 10 年后，30 年后，50 年后这个世界会变成什么样？

蒂姆 • 厄班则首先分析了弱人工智能和强人工智能之间存在的巨大技术挑战，转而又指出，科技发展的加速度规律可以让强人工智能更早实现：“硬件的快速发展和软件的创新是同时发生的，强人工智能可能比我们预期的更早降临，因为：1）指数级增长的开端可能像蜗牛漫步，但是后期会跑的非常快；2）软件的发展可能看起来很缓慢，但是一次顿悟，就能永远改变进步的速度。”

然而，强人工智能一旦到来，人类就必须认真考虑自己的命运问题了，因为从强人工智能“进化”到超人工智能，对机器而言，也许只是几个小时的事情。因为一个可以像人一样学习各种知识的计算机，它的学习速度一定比人快无数倍，它的记忆力一定是过目不忘，它可以从互联网上接触到并牢牢记住的知识一定是这个世界上的全部知识。那么，一个有着和人一样的思考水平的机器，同时有着比人快无数倍的思考速度，以及几乎无限的记忆空间，这台机器在知识理解上能达到什么样的境界？这样的机器几乎肯定比人类所有科学家都厉害！

蒂姆 • 厄班的推理足以让每个读者惊出一身冷汗：“一个人工智能系统花了几十年时间到达了人类脑残智能的水平，而当这个节点发生的时候，电脑对于世界的感知大概和一个四岁小孩一般；而在这节点后一个小时，电脑立马推导出了统一广义相对论和量子力学的物理学理论；而在这之后一个半小时，这个强人工智能变成了超人工智能，智能达到了普通人类的 17 万倍。”

也就是说，一个具备了人类水平认知能力和学习能力的机器，可以借助比人类强大得多的计算资源、网络资源甚至互联网知识库，以及永不疲倦、不需要吃饭睡觉的特点，无休止地学习、迭代下去，并在令人吃惊的极短时间内，完成从强人工智能到超人工智能的跃迁！

那么，超人工智能出现之后呢？比人类聪明好几万倍的机器将会做些什么？机器是不是可以轻易发明足以制服所有人类的超级武器？机器必将超越人类成为这个地球的主宰？机器将把人类变成他们的奴隶或工具，还是会将人类圈养在动物园里供机器“参观”？那个时候，机器真的还需要我们人类吗？

逻辑上，我基本认可蒂姆 • 厄班有关强人工智能一旦出现，就可能迅速转变为超人工智能的判断。而且，一旦超人工智能出现，人类的命运一定是难以预料的，这就像美洲的原始土著根本无法预料科技先进的欧洲殖民者到底会对他们做些什么一样简单。

但是，蒂姆 • 厄班的理论有一个非常关键的前提条件，就是上述有关强人工智能和超人工智能发展的讨论是建立在人类科技总是以加速度形式跃进的基础上的。那么，这个前提条件真的在所有情形下都成立吗？

我觉得，一种更有可能出现的情况是：特定的科技如人工智能，在一段时间的加速发展后，会遇到某些难以逾越的技术瓶颈。

有关计算机芯片性能的摩尔定律（价格不变时，集成电路上可容纳的元器件数目约每隔 18 到 24 个月便会增加一倍，性能也将提升一倍）就是一个技术发展遭遇瓶颈的很好的例子。计算机芯片的处理速度，曾在 1975 年到 2012 年的数十年间保持稳定的增长趋势，却在 2013 年前后显著放缓。2015 年，连提出摩尔定律的高登 • 摩尔（Gordon Moore）本人都说：“我猜我可以看见摩尔定律会在大约十年内失效，但这并不是一件令人吃惊的事。”①

正如原本受摩尔定律左右的芯片性能发展已遭遇技术瓶颈那样，人工智能在从弱人工智能发展到强人工智能的道路上，未必就是一帆风顺的。从技术角度说，弱人工智能与强人工智能之间的鸿沟可能远比我们目前所能想象的要大得多。而且，最重要的是，由于基础科学（如物理学和生

① Gordon Moore: The Man Whose Name Means Progress, http://spectrum.ieee.org/computing/hardware/gordon-moore-the-man-whose-name-means-progress

物学）尚缺乏对人类智慧和意识的精确描述，从弱人工智能发展到强人工智能，其间有很大概率存在着难以在短期内解决的技术难题。

如果蒂姆·厄班所预言的技术加速发展规律无法与人工智能的长期发展趋势相吻合，由这一规律推导出的，超人工智能在可见的近未来即将降临的结论也就难以成立了。

当然，这只是我个人的判断。今天，学者们对超人工智能何时到来的问题众说纷纭。悲观者认为技术加速发展的趋势无法改变，超越人类智能的机器将在不远的将来得以实现，那时的人类将面临生死存亡的重大考验。而乐观主义者则更愿意相信，人工智能在未来相当长的一个历史时期中都只是人类的工具，很难突破超人工智能的门槛。

霍金的忧虑

担忧超人工智能，对人类未来持悲观态度的人有不少。其中，理论物理学家，《时间简史》的作者霍金是最有影响的一个。早在谷歌 AlphaGo 在公众中掀起 AI 热潮之前，霍金就通过媒体告诉大家：“完全人工智能的研发可能意味着人类的末日。”①

作为地球上少数有能力用数学公式精确描述和推导宇宙运行奥秘的人之一，霍金的宇宙观和科技史观无疑是值得重视的。事实上，霍金并不否认，当代蓬勃发展的人工智能技术已经在许多行业发挥着至关重要的作用，但他所真正忧虑的，是机器与人在进化速度上的不对等性。霍金说：“人工智能可以在自身基础上进化，可以一直保持加速度的趋势，不断重新设计自己。而人类，我们的生物进化速度相当有限，无法与之竞争，终将被淘汰。”

此外，霍金同时还担心人工智能普及所导致的人类失业。霍金说，“工厂自动化已经让众多传统制造业工人失业，人工智能的兴起很有可能会让失业潮波及到中产阶级，最后只给人类留下护理、创造和监督工作。”②

基本上，霍金的担忧还是建立在人工智能技术将以加速度的趋势不断增速发展的基础上。如果我们假设这一基础的正确性，那么，霍金的逻辑推论与我们之前谈到的“奇点”理论并没有本质的区别。反之，如果人工智能在未来的发展不一定永远遵循加速度趋势，那么，霍金有关人类终将被淘汰的结论就未必成立。

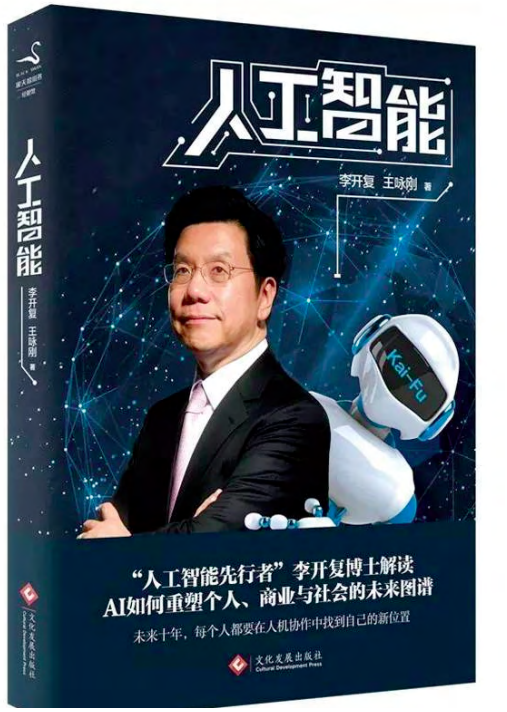
特斯拉与 SpaceX 公司创始人，被誉为“钢铁侠”的埃隆·马斯克（Elon Musk）与霍金有大致相似的担忧。马斯克说，“我们必须非常小心人工智能。如果必须预测我们面临的最大现实威胁，恐怕就是人工智能了。”③

事实上，从行动上看，霍金和马斯克并不是简单的悲观主义者，他们在警告世人提防人工智能威胁的同时，也在积极行动，试图为人类找出应对未来潜在威胁的对策。马斯克说，“我越来越倾向于认为，也许在国家层面或国际层面，必须有一种规范的监管机制，来保证我们不会在这方面做任何蠢事。”

除了呼吁建立监管机制外，马斯克还与萨姆·奥尔特曼（Sam Altman）一起创立了非盈利的科研公司 OpenAI。谈到创立 OpenAI 的初衷，马斯克说，“为了保证一个美好的未来，我们最需要做什么？我们可以冷眼旁观，我们也可以鼓励立法监管，或者，我们也可以将那些特别关心如何用安全的、对人类有益的方式来开发 AI 的人合理地组织起来研发 AI。”④

如果说这个世界上还有几家纯粹理想主义的公司的话，OpenAI 一定算一个。OpenAI 一面聚集了一批 AI 领域的顶尖高手，研发最前沿的 AI 技术（主要是强化学习和无监督学习技术），甚至探索实现强人工智能的可能性，一面反复强调

① Stephen Hawking warns artificial intelligence could end mankind, http://www.bbc.com/news/technology-30290540
② 霍金：自动化和人工智能将让中产阶级大面积失业，http://tech.qq.com/a/20161203/002359.htm
③ Elon Musk: artificial intelligence is our biggest existential threat, https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat
④ Silicon Valley investors to bankroll artificial-intelligence center, http://www.seattletimes.com/business/technology/silicon-valley-investors-to-bankroll-artificial-intelligence-center/



自己的使命是研发“安全的”人工智能，通过实践来探寻将人工智能技术的潜在威胁降至最低的方法。

马斯克和奥尔特曼的 OpenAI 看上去是在做一件自相矛盾的事情：既积极地研发人工智能甚至是强人工智能，又希望将人工智能关在道德或制度的“牢笼”里，让 AI 难以威胁人类。事实上，目前 OpenAI 所开展的工作，和其他人工智能科研机构所做的并没有本质的不同。据说，OpenAI 的研究总监伊尔亚·苏茨克维（Ilya Sutskever）表示，OpenAI 最重要的目标，就是发表有影响力的文章。[①]或许，马斯克和奥尔特曼的意思是说，既然奇点来临无法避免，那不如积极投入，至少，当威胁来临，我们对威胁本身的理解会更加深刻。

2017 年初，霍金和马斯克均表示，为了防止人工智能威胁人类，他们支持加州阿西洛马（Asilomar）会议通过的 23 条基本原则。[②]这 23 条基本原则涵盖了三个范畴：1）科研问题，2）伦理和价值观，3）长期问题。

阿西洛马 23 条基本原则像科幻大师阿西莫夫笔下著名的“机器人三定律”一样，从方法、特征、伦理、道德等多方面，限定未来的人工智能可以做什么，不可以做什么。例如，有关人工智能相关的伦理和价值观，其中几条原则是这样规定的：[③]

安全性：人工智能系统应当在整个生命周期内确保安全性，还要针对这项技术的可行性以及适用的领域进行验证。

价值观一致性：需要确保高度自动化的人工智能系统在运行过程中秉承的目标和采取的行动，都符合人类的价值观。

由人类控制：人类应当有权选择是否及如何由人工智能系统制定决策，以便完成人类选择的目标。

非破坏性：通过控制高度先进的人工智能系统获得的权力，应当尊重和提升一个健康的社会赖以维继的社会和公民进程，而不是破坏这些进程。

应当说，在担忧未来人工智能威胁的人中，霍金和马斯克还是一直抱有一种非常积极的态度的。他们一方面基于自己的逻辑判断，相信人类未来面临机器威胁的可能性非常大，另一方面又利用自己的影响力，积极采取行动，尽可能将人工智能置于安全、友好的界限内。从这个角度讲，霍金和马斯克至少比那些盲目的悲观主义者，或因未来的不确定性而丧失勇气的怯懦者强很多很多倍。

（本文根据原文整理，文字有删减，已经作者确认。）

（责任编辑：钟宇欢）

① 如何评价 Elon Musk 启动的 OpenAI 项目？ https://www.zhihu.com/question/38441799
② ASILOMAR AI PRINCIPLES, https://futureoflife.org/ai-principles/
③ 人工智能的 23 条“军规”，马斯克、霍金等联合背书，http://tech.qq.com/a/20170207/031641.htm