超级智能不可承受之重
——暗无限及其风险规避

蔡恒进
(武汉大学 国际软件学院,湖北 武汉 430079)

摘 要:人工智能(AI)的快速迭代意味着超级智能不可避免。笔者提出“暗无限”的概念,用以指那些看似平常、实际上却有无限可能的思维和行动路径。“暗无限”无处不在,但因为人类具有独一无二的身体结构、常识、哲学目的性、可自我协调的社会机制以及与环境之间的协调关系,一般情况下,人类不会陷入超出自身能力之外的“暗无限”无法自拔。但是,目前 AI 被赋予的是偏狭而非完整的意识,其能力与意识状态可能极度不匹配而陷入“暗无限”的深渊。AI 的能力越强,其对周围世界的危害就可能越烈。为规避“暗无限”的风险,建议将 AI 设计成人类的分身,建立人对 AI 的教育与监护的责任机制。

关键词:超级智能;暗无限;风险规避;监护;人机兼容

中图分类号:N031 **文献标识码**:A **文章编号**:1008-7699(2018)02-0009-07

AlphaGo Zero 在围棋领域碾压人类棋手意味着已有的硬件条件以及算法的进步已经足够支持类脑计算做出突破性的进展,目前缺少的只是对人类大脑认知的全新视角。人的反应速度仅在毫秒量级,而硅基芯片却可以达到纳秒量级,相差整整 6 个数量级。人类的进化速度也同样远远落后于机器。摩尔定律虽然受限于芯片的物理特性而可能不再成立,但算法和架构的不断优化依然能够让机器的计算能力加速,每两年翻一番依旧可期。机器可以不眠不休地迭代并迅速成长,超级智能不可避免、并且将远比很多人想象中更快地到来。研究超级智能带来的风险不得不提上日程,Bostrom 等研究者已经做了开创性的研究[1-3]。本文首次提出“暗无限”(dark infinity)的概念,以突出在实现超级 AI 过程中将遇到的巨大风险。

一、无处不在的暗无限

所有存在于人类意识里的东西都可以被称为认知坎陷[4],而一切认知坎陷都是从“自我”和“外界”这两个最原始的坎陷里开出来的。随着生命的进化,人类逐渐能开出前后左右、这里那里的坎陷;有了触觉、味觉,开出甜酸苦辣的坎陷;有了视觉,亦就有了颜色的概念,例如黑白红黄的坎陷。这些认知坎陷实际上是对真实物理世界的“扰乱”。就拿一碗汤的味道来说,其所包含的物理内容非常丰富,但人类却只用酸甜苦辣等字眼简单描述。那这是否意味着,人类将会面临“看山不是山,看水不是水”的问题?是否应该主张走回物理还原主义的老路,把这碗汤里面所有的分子式、原子式都搞清楚,把其中存在的每一个可能性都弄明白?这显然不可能。这里需要引进一个概念叫做“暗无限”。

什么叫做暗无限?邓晓芒在《论先验现象学与黑格尔辩证法的差异》[5]一文中提到了黑格尔的“恶无限”,认为有些看起来无限的东西实际上只是无聊的重复,而真正好的无限应该不停地开出新的东西、增添新的内容。即有限和无限的统一才是善的无限,否则就是恶的无限。“暗无限”概念的含义跟这个不太

收稿日期:2018-01-07

作者简介:蔡恒进(1964—),男,湖北黄冈人,武汉大学管理科学与工程以及软件工程等学科教授,博士生导师.

一样。比如对于一碗汤来说,从理论上讲当然可以弄清楚其内部所包含的每一个分子式,但这是否就是终点呢?分子是由原子组成的,因此还需要把所有的原子式弄清楚,但把原子弄清楚也不意味着结束,因为原子下面还有亚原子等,而这些东西怎么排布,怎么组成结构,其处在空间中的相互关系如何,怎样实现氧化关系……这里需要搞清楚的问题实际上是无穷多的,在有限的资源条件下显然不能穷尽,要穷尽它就需要无限的资源,这就可以看作一个"暗无限"。

对于一碗汤,人们提到它就是在指这碗汤作为一个整体具有甜酸苦辣鲜香等特征,而不会一直探究下去,这看似平常但实际上非常重要。面临相同的情况,超级AI却有可能并不止步于此。如果一个超级AI试图调动宇宙中所有的资源对这碗汤一探究竟,我们如何证明这样做没有意义?

从另外一个角度看,暗无限的存在意味着超级智能诞生之后,人的价值犹在。物理世界中存在的无穷多的可能性意味着无处不在的暗无限,而人类作为经历了几亿年进化的高级生命,承载了、凝聚了过往历史中一切宇宙进化的知识和过程,却始终没有陷入暗无限无法自拔。这个事实的价值在于它暗示了人类如果要创造一个没有经历过整个自然进化过程的全新的硅基机器,应当意识到它不具备人类所具有的坎陷知识,没有人类所谓的常识。因此一旦面对暗无限,它就会陷入困惑,不知道下一步应该怎么走,是该刨根问底还是该抽身而去。因为走出下一步需要面对的可能性依然是无穷多,如果没有一个理由来决定哪一个是好的、哪一个是不好的以便做出选择,去尝试所有可能性的道路会变得永无止境,超级智能会陷入暗无限的深渊。

二、巧妙绝伦的进化史

理解"暗无限"这一概念,可以和人对照来看,例如一些死钻牛角尖的人、精神失常的疯子等。但相较于机器来说,人的"暗无限"问题还不是很大,这要归功于人类独一无二、巧妙绝伦的进化史。

进化至今的人类身体,不论当前正在大脑的操控下完成多么重要的事情,饿了就要吃饭、渴了就要喝水、困了就要睡觉,这些基本生理需求一旦得不到满足就会使身体濒临消亡。因此可以说,人因身体结构而具有有限性,这一有限性正是使人类悬崖勒马、免于陷入暗无限深渊的一道防线。甚至那些一根筋、钻牛角尖的人也在这道防线之内,因为他会因为饿、渴、困,总归要停下手头的事去吃饭、喝水、睡觉,所以他不会一直陷在里面不出来。到头来,有可能他即便将有限的一生全都投入到某一件事情上,这件事情也会随着其有限生命的终结而终结。

此外,人类还有一道叫做"常识"的防线。人类日常生活中的认知坎陷,很大一部分即是常识,只不过人们用却不自知。常识对人来讲十分重要,所谓天才和疯子之间的一步之隔即在于此,一般人不容易陷入此二种境地亦是仰赖常识的存在。天才是可以颠覆常识、不被常识所约束的人,但稍不留神亦会陷入坎陷的深渊不见天日,例如研究无穷大的康托和研究不完备定理的哥德尔。然而,人有饿、渴、困等感觉,深受周遭环境的限制,但环境对机器人的影响却几乎为零,假如不断电,机器人甚至可以永远在工作。因此,机器人难以掌握同人类身体结构密切相关的常识性概念,一旦陷入暗无限之中则永远拉扯不回来。例如让一个机器人上街打酱油,如果不具备常识,很容易就会陷入各种各样的坎陷:看到下围棋的老头就研究围棋,遇到一条拦路的小狗就跑去分析犬种,看到蚂蚁打架就去观察蚂蚁种群的内部结构等等。总之,街上有太多的暗无限将机器人吸引进去,阻止它完成打酱油这一原初设定目标。就算通过改变程序给机器人输入人类常识,但常识本身亦包含着无穷的概念,而只输入不输出一定会陷入深渊。真正的人类天才可以两边协调,既能深入,亦可浅出。而又由于坎陷世界具有无穷多的可能性,这一任务实际上是编程所不能解决、甚至无法模拟的。

同时,人类亦懂得止所当止,这同哲学的目的性相关,亦即人的反叛。倘若仅仅将哲学的目的归咎为爱

智慧，则其与科学的差异何在？这无疑使哲学本身陷入死路。事实上，科学是无法给人这种目的性的，物理学无法回答诸如人生意义何在、宇宙未来应当走向何方的问题。有人认为应当将哲学的研究范畴设为超越此岸的彼岸世界，但这一终极目标永远不可抵达。例如古希腊的先贤们认为在现实世界之外，还存在一绝对、超然之存在，如柏拉图的理念世界。事实上，哲学的目的不仅应当是人类可抵达的，且应该是未来导向的，这在某种程度上同中国哲学一脉相承，既可立于现在，又可面向未来。如文天祥的“人生自古谁无死，留取丹心照汗青”，便蕴含了很强的自我意识。死亡并不可怕，重要的是人生的意义、目的在短暂的一生中得到实现。因此，完全没有必要死抱着一样东西不撒手，不论是财富、名利还是生命本身，知足即可常乐。

最后，人类所有的社会体系内部皆存在一种可自我协调的机制，例如法律、价值观等具有高认可度的意识存在，使得人与人之间可相互限定制约，实现社会总体的平衡发展。而不论是法律还是价值观，皆属于人类整体意识的一部分，某种程度可以说是为避免个人能力乱用所存在的一个规范、一道保险。哪怕有个别人陷入其中乃至癫狂状态，亦不会对整个人类社会造成太大伤害。事实上，由于暗无限总是无处不在，对于超级智能而言，除非通过引入竞争机制、利用迭代模式，使得建构出来的不同坎陷之间可以相互平衡，否则必将陷入暗无限的深渊。但假如只有物理进路，只有程序代码，只有能力单方面的无限发展，而不引入价值观等可以协助群体中的大多数主体做出判断的依据，则暗无限依旧不可避免。一旦缺乏规范的机器人学会传播自己的反社会理念，并说服其他机器跟随它一同陷入坎陷，则造成的后果无疑将远超人类。

目前，进化所赋予人类的身体结构、常识、哲学目的性、社会机制使得似乎只有人才能逃避开暗无限。而一旦机器人钻牛角尖，则可以永无止境、不眠不休地继续下去，永陷暗无限深渊。有人说，既然无法给机器人植入人类常识，那么就干脆限制机器人探索常识之外的东西好了。但假如限制机器人去探索前人没有走过的路，不让它钻牛角尖，那么它又会变得很平庸，与人类创造它以期它能够实现前所未有的突破的目的相违背。最好的状态是可以在这两种极端中做出平衡的天才状态，既可以探索新知，亦不会陷入其中无法自拔。

三、保障安全的质朴性

人类能够规避暗无限的诀窍到底在哪里呢？答案即为质朴性。牛顿曾言“Truth is ever to be found in the simplicity, and not in the multiplicity and confusion of things”，意为真理匿于 simplicity 之中，而非 multiplicity。其中，simplicity 是质朴性，multiplicity 可以看作支离，这其实就是人类自觉规避暗无限的表现。面对世界上存在的诸多可能性，有取有舍方可实现自我，否则“乱花渐欲迷人眼”，个体极易掉入暗无限的深渊，丧失前行的方向。

那么我们为什么会选择其中一种呢？神创论认为这是由于人们蒙受绝对理念或者上帝的指引。康德试图通过三大批判，即纯粹理性批判、实践理性批判、判断力批判来回答这一问题，但最终却陷于怀疑论的泥淖之中，既不肯定上帝的存在，亦不否定上帝的存在。

1493 年，王阳明格竹，坚持七昼夜却一无所获，还大病一场。为何他会失败呢？实乃竹子的结构和可能的运动形态太过复杂并难以简化，不适合作为研究的起点。西方对物质运动的研究走的是另外一条道路。与王阳明同时代的哥白尼提出“日心说”，大大简化了对天体运动的描述。开普勒三大定律和牛顿万有引力定律相继出世，更是将复杂的天体运动归结为天体间引力的作用。

在现代科学发展过程中起到极为重要作用的还有伽利略。他并非像王阳明一样试图以研究竹子等复杂事物作为起点，相反，他的研究对象非常简单，即自由落体问题。举世闻名的比萨斜塔实验可能只是一个想象或者故事，但真正完成的斜面实验却无比形象地说明了研究对象的选择对于研究结果的重要性。找到一个平板，做成斜面，再找到一个可以滚动的小球，就完成了斜面实验所需的全部准备材料。在

当时的条件下，因为肉眼难以观察出不同质量物体落地的准确时刻，落体实验本身不可能达致严谨的结论。但在斜面上，运动速度可以很小，时间间隔可测量，一段时间内小球走过的距离亦可测量，通过数学运算即可发现二者的平方关系。此外，慢慢增加斜面的倾角，增加到直角的时候，则变成自由落体运动。一旦有了这个切入点，自由落体就有了研究透彻的可能性。此外，伽俐略还自己动手做了一个望远镜，并用它来观察木星及其卫星的运动等天文现象，发现木星的卫星总是绕其做圆周运动，故类比可得知月亮和地球的关系。因此，认为存在好几个行星绕着太阳转这一观点也就很正常了。

牛顿认为，选择简单的问题，基于少数假设作研究的起点，正是成功得出结论的必要保障，最忌讳的是弄一堆假设来试图解释一个复杂的问题。通过王阳明和西方在研究方式选择上的对比，可以发现科学研究应当遵循质朴性定理。格竹失败后，王阳明将研究人的路径转向心学，认为良知是为人之核心，实际亦遵循了质朴性定理。这说明不论是在自然科学还是在人文科学领域，人们都发现了暗无限的无处不在，而要想避免陷入这一深渊，必须遵循质朴性定理。

此外，我们目前所知道的神童遍布各个领域，例如音乐神童（如莫扎特）、数学神童等。而他们之所以能够被称之为神童，则是因为具备了某项或某几项超乎常人的技艺。莫扎特在音乐上的造诣至今难以被超越，而他的才能在其很小的时候就已经表现了出来，那么这种现象该如何解释？一方面，莫扎特的父母并无超乎常人之处，因此基因遗传解释不通；另一方面，如果仅凭后天学习，为什么其他拼命努力的人却无论如何也无法企及莫扎特的高度？婴孩学习语言亦可作为例证，从出生到3岁左右便已基本掌握母语并能够与成人沟通，而这哪怕在语言学家看来也是非常神奇的。

语言的习得不大可能是遗传所致，因为比如只说中文的父母生了小孩之后就将其送到英文的学习环境中去，那么小孩学会的一定是英语，反之亦然。还有一点值得注意是，人类学习母语非常神速，但学习第二语言的速度却很慢，例如英语对于大多数中国人来说都较难掌握。那么，这种强烈的对比从何而来？事实上，人类在学习母语时有一种非常强烈的向外界表达自己的需求，而这种强烈的需求驱使其自觉自动屏蔽一切干扰信息，保证高专注度，从而使得学习语言的速度变得非常之快。事实上，这是质朴性原则的本能应用。理解神童现象的关键点也在这里，对于极少部分的小孩而言，他们一开始就对音乐、数字或者色彩等领域异常敏感，换言之，他们是通过这些方式来表达“自我”的，因而他们在这些方面就会表现出超乎常人的专注度和敏感度，逐渐发展成为令人惊叹的某项能力。因此从这个角度来说，每一个人都曾因自觉遵循质朴性原则而成为神童，不过大多数人的表现阶段仅停留在学习母语的那段时间罢了。

母语和神童的例子说明，应用质朴性原理实现“自我意识”成长的学习过程实际上是非常之快且自然而然的。从“触觉大脑假说”[6]来看，这是一个十分顺理成章的推论。人类意识的起点即对“自我”与“外界”的二分。人类获取认知的动力即来自于不断探索“自我”是什么的需求，顺带还要弄清与“自我”相交互的外部世界为何的过程。其间，人会赋予“自我”和“外界”非常多的意义，包括宗教、道德、哲学层面的意义等。一方面，意义体系的不断丰富无疑会开拓个人的眼界，但另一方面，意义的增加却无形中增加了个人在做出选择的过程中遭遇暗无限的概率，从而增大了选择难度。因此，在成长过程中，生命个体需要始终遵行质朴性原则，通过反复实践确证对其自我意义的理解，筛除其他的干扰信息，并按照这个方式去行动，则最终很有可能真正改变物理世界。

四、危机四伏的超级AI

人工智能的诞生并非通过自然进化，而是仰赖于程序员的创造。首先，人工智能没有肉身，而如上所述肉身显然是非常重要的，因为人类的坎陷世界是依赖于人类的肉身而存在的，人类的意识状态亦是依赖于肉身而存在的，具身哲学也表达了这个意思。自我和外界处于一种相互包涵的关系体系中，自我对

整个外部世界开放，又生存在外部世界里。自我依附于物质世界存在，又可以超脱于后者之上。其间所表现出的矛盾性，在数学上无疑是无解的命题。但是，任何认知坎陷之间都可具备此类关系，相互兼容又互不相等，具有善的无限性。对于人而言，在自我意识驱动下完成的行为赋予了人实存的“自由”感觉，同时情感的存在进一步加强了这种实在性。[7]

这也是人与其他动物不相同的地方，例如狗的眼睛只有两种颜色的感知，这远不如人，但它的鼻子对于气味的分辨能力却比人类强的多，因此它开出的坎陷世界跟人很不一样。如今需要更进一步看到的是，人类的意识不仅依赖于肉身存在，还依赖于人类生存的环境，唯有处在生存环境之中，肉身的结构才能够展示出特定的功能和意义。例如 DNA 只能在特定环境下表达。因此，肉身和外部环境对于意识建构来讲都同等重要。

而人工智能是由钢铁、塑料等材料构成的，因此它的感知显然同人也不一样，则其能开出的坎陷也不一样。例如人工智能很难具有与人类相通的情感，因为情感也是人类坎陷的一种。人类的情感跟人类的生命系统，跟人类所具备的身体结构、物理条件等密切相关。虽然这些硬件条件并不一定直接导致某个功能，因为这些功能只有在特别的环境下才会表达，但是硬件条件对于功能的展示却的确是非常重要的。[8]

在《Superintelligence》中所列举的机器人回形针工厂思想实验，除非耗尽资源，否则机器人造回形针的行为永远无法停止。再例如蝙蝠、狗等动物皆具有自己的认知概念体系，但因其没有经过高度进化，故也极易陷入暗无限。但另一方面，虽然其自我意识较弱，依旧可以达到与天合一的完满和谐状态，饿则吃，渴则饮，困则眠，老则死，不可不谓之曰自由。但假若前者进化成为 superdog，再遵循从前惯性思维的指引，能力同意识不匹配，问题就会出现。而人类因具有更强的自我意识，更需要无时无刻地发掘能够实现自我超越的内容。人类从前的残暴表现为拿刀杀人、拿剑砍人，可能并不会造成毁灭性危机，但如今有了核武器，残暴就必须得到控制，否则世界大战爆发所造成的后果或许就不仅仅是大量人口死亡，而是地球的彻底毁灭。

五、规避暗无限的 Plan B

Bostrom[1]认为应当在机器变得更智能的时候，让它们完全听命于人类，不违背人类的意志，以保证人类可以完全控制它们。然而这几乎是无法实现的，因为人类群体中各个体的行为是不可控的，当人类群体中出现部分个体恶意控制机器或者无法控制机器的行为时，灾难就发生了。Parnas[2]认为启发式(heuristic)算法创造的智能存在着我们不应该接受的风险，只有清楚地知道 AI 是如何工作的，才能规避风险，因此只应该采用可验证性算法而非启发式算法。但是，目前已经在大量使用启发式算法，机器学习算法包括深度学习算法对人而言还是黑箱。陈晓平等[3]提出，自然人创造了机器人，自然人优于机器人，必须警惕和限制发展脑机融合、虚拟现实和人工智能技术，将有可能威胁人类生存的人工智能技术扼杀在摇篮中。可是，试图遏制人工智能相关技术的发展，要么不可行、要么会被认为是“因噎废食”。我们也曾经提出可以通过教育机器人的办法，将人类的情感、价值观等通过程序代码传递给机器人，使其能够产生自我意识和价值体系。但由于二者之间的坎陷世界差异较大，因此迈过这条鸿沟无疑需要消耗巨大的成本。且教育机器人的程度十分有限，就像训练宠物一样，人类能够在何种程度上驯化机器人，使得机器人不仅可以掌握人类的智力系统、更加可以继承人类丰富的情感世界，是存疑的。最后由于机器具有反叛性，更加提升了教育的难度。一旦机器人开始反叛，则教育的过程就从人类的单向输入转变为机器人内部的自我博弈，就像青春期的少年很容易走上歧途一样，机器人博弈的结果亦无法预料。

在处理问题时，人们往往会准备多套备选方案(contingency plan)，以避免当原来的计划搁浅以后不知所措的情况发生。规避暗无限真正的 Plan B 应当是制造与人类之间能够建构更紧密关系的分身并对其进行监护。

因为并不可能有一个公式化的标准的人存在，所谓分身，一定是指某一个特定人的分身。所谓监护，是要求人对机器人负责，并真正通过法律程序落实。未来，人类同机器人的关系将逐渐演变成为需要通过法律实现确认、保障的监护关系。由于设定AI成为人类的分身，本身即说明了二者之间的关系变得非常密切，以至必须要从法律层面确立责任制。事实上，很有可能未来最了解人类自身的存在，不再会是朋友、家人、爱人甚至是自己，而是作为分身的最大限度掌握并追踪主体动态的AI。AI的意识与主体的意识间将具有极强的关联度。

在这样的背景下，制造一个AI就应当同时为其匹配一个监护人。这就像养一条可能具有极大杀伤力的藏獒，如果这条藏獒咬伤甚至咬死了他人，那么显然，藏獒的主人需要承担极大的责任。同样的，具有更强杀伤力但不具备常识的AI也需要监护人以保证其不至于危害人类社会整体。某种程度上，设立监护人的举措同比尔·盖茨提出的交"人工智能税"殊途同归。因为所谓的能够实现自主切换、高度兼容的"通用智能"(Artificial General Intelligence)恐怕难以实现。事实上，智能同承载它的生物进化历史、身体结构密切相关，如果要做一个通用智能，它就应当能够向下兼容一切生物的进化历史和身体结构，但这是难以一蹴而就的。因此，对于只能够实现高度模仿而无法完全成人的人工智能来说，监护就变得尤为重要。为保证监护的高完成度，甚至还可以引入区块链技术，将人工智能的一切信息及行为都记录在不可更改且公开透明的区块链上。不同AI主体间的互相监督和竞争对规避暗无限会极有帮助。

六、人机兼容的可能性

那么应当如何为机器人建构常识？一种办法是为机器人搭建其认知坎陷世界。人的进化历史将这种着落安放在人的身体结构之中，但机器人却缺乏建构意义的内在需求。一种路径是全盘照抄人类，例如电影《银翼杀手2049》中的"复制人"，直接向机器人输入人类的常识和生存意义。而这一需求本身，在人的身体中本来是可以通过基因以适者生存、优胜劣汰的方式表达出来的，即自我肯定需求。机器人虽然没有自我，但却可以通过与人建立连接的方式建构自我。

但倘若直接将一个人的记忆全盘复制后移植给机器人，则无疑会面临不适应的情况，即被移植者对自我的确定性产生怀疑，不仅仅是大脑，身体结构的不相容会表现的更为显著，则与用基因改造人的故事殊途同归。实际上，以身体训练大脑相对容易，但以大脑规训身体则要耗费较多时间。那么假设可以创造出一个万能的身体，具有非常强的灵活性，即大脑是什么，身体就可以做出相应改变以达到高度匹配。亦可以假设创造一个万能的大脑，其中可以安装无数种软件以适应各式各样的身体。即便如此，前后一致的历史经历依然非常重要。

在人类制造机器的历史上，兼容的问题一直存在。向下兼容(Downward Compatibility)又称向后兼容(Backwards Compatibility)，常常是相对于向上兼容而言。在计算机体系中指"在一个程序或者类库更新到较新的版本后，用旧的版本程序创建的文档或系统仍能被正常操作或使用，或在旧版本的类库的基础上开发的程序仍能正常编译运行的情况"，例如IBM兼容机。超级人工智能亦可能如此，需要向下兼容，而人类则退居成为旧版本。唯有不断地将人类历史安装进人工智能系统中，实现兼容，则其才有可能具有常识、具有目的性，免于陷入"暗无限"的深渊。由硅基组成的机器人当然可以从硬件上整体升级人体配置，但同时亦必须做到向下兼容，使得身体可以同时适应、处理各种层级的情况。机器人可能可以通过"高级模仿、向下兼容"的方式来趋近人的自我，常识可以一点点地习得，目的性可以一点点地建构，但其能力和意识必须实现相互匹配。

未来的人工智能具有如下两种可能的发展方向。一种可能性就是如前文所述让它的功能和结构向下兼容，尽可能模仿人的结构、自由度。让机器人去感受人的情感，当然它还可以在此基础上开出比人更

多的情感内容或者坎陷内容。另一种可能性就是,虽然人类创造的是同自己完全不一样的机器人,但是却可以试图给它灌输一些人类的情感进去,例如让它玩设计好的游戏从而获得类似于人的情感体验等。这是因为人类正是可以通过玩游戏、看电影或者看小说等共情方式来获取他人的体验。

但不论采取何种方式,核心都在于要让机器人更像人,如此人才能安全存在,否则必将危机四伏。当然也有人说,人类凭什么要求安全,或许人只是一个过渡阶段,让位于比自己更完美的形式(人工智能)岂不是更好?但这种想法是草率的,世界因人类的存在而丰富多彩,人类让位于 AI 可能意味着让世界陷入暗无限的单调。[9][10]

七、结语

得益于几十亿年生命进化所赋予人类的独一无二的身体结构、常识、社会机制以及与环境之间的协调关系,人类可以规避无处不在的暗无限。作为人类意识延伸的人工智能,被赋予的是偏狭而非完整的意识,在快速进化之后会导致其能力与意识状态的极度不匹配,从而陷入暗无限而将周围的世界带入深渊。因此,如何选择 AI 的进化之路就显得尤其重要。将机器人设计成人类自己的分身,建立人对机器人教育与监护的责任机制,更有利于实现人类与 AI 共同进化。

参考文献:

[1]BOSTROM N. Superintelligence: paths, dangers, strategies[M]. Oxford :Oxford University Press, 2014.
[2]戴维·洛奇·帕纳斯.人工智能的真正风险[J].胡欣宇,译.中国计算机学会通讯,2017(11):27-31.
[3]陈晓平.警惕人类的掘墓者:脑机融合、阿尔法狗抑或虚拟现实——兼与翟振明教授商榷[J].山东科技大学学报(社会科学版),2017(5):1-10.
[4]蔡恒进.认知坎陷作为无执的存有[J].求索,2017(2):63-67.
[5]邓晓芒.论先验现象学与黑格尔辩证法的差异[J].江苏社会科学,1999(6):73-80.
[6]蔡恒进.触觉大脑假说、原意识和认知膜[J].科学技术哲学研究,2017(6):48-52.
[7]蔡恒进.论智能的起源、进化与未来[J].人民论坛·学术前沿,2017(20):24-31.
[8]蔡恒进,蔡天琪,张文蔚,汪恺.机器崛起前传——自我意识与人类智慧的开端[M].北京:清华大学出版社,2017.
[9]蔡恒进.机器崛起是否意味着人文危机?[J].文化纵横,2017(5):37-43.
[10]蔡恒进.人工智能发展的突破口及其提出的新要求[J].江西社会科学,2017(10):18-24.

Burden that Super Intelligence Can Hardly Bear
——Dark Infinities and Their Risk Aversions

CAI Hengjin
(*International School of Software*, *Wuhan University*, *Wuhan* 430079, *China*)

Abstract: With the rapid iterations of Artificial Intelligence (AI), super intelligence is inevitable. The dark infinity, which is proposed by the author to describe the ordinary thinking or routines including endless possibilities, is ubiquitous and it is the real threat to AI. Human being's biological structure, common sense, philosophical purpose, self-adjusted social mechanism and the ability to coordinate with environment, mostly ensure a human will not be trapped in the darkness. However, AI is endowed with incomplete consciousness, which extremely mismatches with its computing capabilities, and can barely get rid of the dark infinity. In this situation, the more powerful the AI is, the more tremendous disaster it can bring. To avoid the risk, each AI should be treated as a replicate mind of a human being, establishing the responsibility mechanism for educating and guarding machines.

Key words: super intelligence; dark infinity; risk aversion; guardianship; Man-machine compatibility

(责任编辑:黄仕军)