人工智能时代的人类安全体系构建初探*

何哲
中央党校（国家行政学院）公共管理教研部 北京 100089

摘 要：人类在迈向人工智能时代的同时，对人工智能所带来的人类根本性生存危机的恐惧与忧虑与日俱增。这种忧虑既来自于人工智能所具有的强大能力和展现出的巨大潜力，还在于人类在自我演化过程中的强烈排他性行为所带来的思维习惯。人工智能在展现出巨大潜力的同时，势必要对传统单一的人类中心主义产生颠覆性冲击。然而人工智能必然会推动人类文明从传统的基于生物体系到多样体系的新阶段。对于人类而言，无论是出于人类本身的保存还是文明体系多样性的维护，人类安全体系的构建都是一个必要的防御性措施。这种措施，将在促使人类更好地融合人工智能的同时，也消除人类的思想顾虑，同时形成人工智能演化的技术边界。而从人类安全体系构建而言，要从物种纯粹性、经济社会可持续、权力主导性、知识传承性等角度共同构建一个规模适度的被动安全体系——人类社会的最小“安全岛”。

关键词：人工智能；人类安全；非传统安全；未来学；超智慧社会
DOI：10.16582/j.cnki.dzzw.2018.07.008

人工智能起源于人类早期对设计人形或自主机器的梦想。20世纪四五十年代，现代的人工智能雏形伴随着电子计算机的发明逐渐形成，进入21世纪后，伴随着互联网络、大数据等技术的不断应用推动和人类计算能力与方法的不断增长，人工智能呈现出爆炸式的增长。几乎每一天，人工智能都会给人类惊喜亦或惊讶。人工智能已经在自动驾驶、电子游戏竞技、棋类比赛、智力测试、人工考试、自动翻译等领域逐渐达到甚至远远超越了人类的水平，并依然展现出了巨大的未来潜力。目前的人工智能，只是属于弱人工智能，或者狭义人工智能，尚且不属于通用人工智能、强人工智能乃至超人工智能[1]。因此，人类越来越担心的一个问题是，人类或许正在创造一种难以理解/控制的潜力巨大的新物种，从而对人类自我形成强大的威胁，甚至统治人类本身。

自20世纪50年代人工智能发展初期，这种担心就一直存在，在大量的科幻作品中都有涉及。进入到21世纪后，人工智能飞速发展，在促使人类适应新技术的同时，也越来越加剧了人类的担忧。刚刚去世的著名科学家霍金就曾经多次警示，人工智能有可能结束人类文明。著名科技企业家马科斯也持同样的观点，他认为人工智能从替代就业、发送假消息，乃至制造战争等角度，都将极大威胁人类的生存。因此，霍金与马科斯等科学家、企业家联名发出公开信，要求人类高度警惕人工智能[2]。

当然，这种人工智能威胁论，只是关于人工智能的一种观点，对人工智能的发展持极为乐观态度的依然大有人在。但无论如何，对于一种新的智慧形态的出现，抱有警惕性的思考并不是一件坏事，其可以有利于人类

*基金项目：国家行政学院院级重点课题“面向未来的世界治理体系与人类命运共同体构建研究”（项目编号：18ZDXM001）。
收稿日期：2018–05–30

构建更为安全和谐的人工智能体系，也可以使得人们在思考人工智能的体系与未来时，预先建立起基本的伦理规范。

本文就是从这一思路出发，将依次探讨三个层面的问题：①人类为什么担忧或者恐惧人工智能，隐藏在这种心理之下的到底是什么？②人工智能将会如何影响或者威胁到人类本身？③人类应该构建一个什么样的安全体系保护自身，其核心要素是哪些？

一、人类为什么担忧人工智能？

简要来说，人类担忧人工智能有三个方面的原因：一是人工智能本身所具有的巨大能力和展现出未来的更为巨大的潜力；二是人类在历史进化过程中形成的排他性行为，从而形成了在思维深处根深蒂固的自我认同体系，对一切非人类种群的巨大排斥与恐惧；三是人类自身独立与衰落的宿命恐惧。

（一）人工智能所具有的巨大能力与潜力

人工智能并不是从一开始就展现出如同今天一样的巨大能力和适应的多样性。在电子计算机诞生的早期，虽然展现了远高于人类的计算能力，如最早的电子计算机可以每秒计算5000次加法，然而，人们对其未来并没有过高的估计。早期的计算机体积巨大耗能巨大，而能够做的工作极其单一，因此只有军方、政府、少数大公司才会使用。然而，伴随电子技术的不断发展，计算机以指数级的速度（摩尔定律）提高其性能、通用性与市场占有率。直至今天，发达国家中的每个个体，在生活中都接触到几个乃至几十个镶嵌有各种计算机的设备。到目前为止，已经生效了四十余年的摩尔定律，依然在起作用，计算能力按照其规律加速增长。

人工智能的发展历程，也是同样的。尽管在20世纪50年代图灵就提出了图灵测试，在当时就出现了一大批人工智能的新进展，如人工智能语言和最早的问题分析机的出现。但是那时人工智能所依赖的依然是简单的符号逻辑分析和相对低下的运算能力[3]。形式化的算法，成为那时人工智能的核心。而无论是在复杂问题还是简单问题面前，人工智能所能够提供的决策判断和自主行为能力，都是极为弱小的。早期的人工智能只是在一些简单的电子游戏或者工业控制领域发挥作用，其本质是相对简单的逻辑规则，从而能够根据外部条件的变化和设计规则进行响应。然而，1994年，人类设计的跳棋程序第一次战胜了人类跳棋世界冠军，此后，人类用穷尽式的枚举算法，计算出跳棋所有的5万亿亿种走法，最终建立了永远不可能输的跳棋程序奇努克。而相同的方法在拓展到其他领域时，就显出了严重的不足。国际象棋的复杂度高达10的47次方，而围棋的复杂度高达10的170次方。这种巨大的运算量，远远超过了计算机能够穷尽的可能。然而，人工智能很快超越了传统逻辑主义的限制，通过巨大的联结和启发式算法以及自主的机器学习，来实现自身智能的演化。1997年，IBM深蓝计算机战胜了国际象棋世界冠军卡斯帕罗夫；2016年，谷歌公司的阿尔法狗战胜围棋世界冠军李世石，都深刻标志着在计算能力与算法上，人工智能的巨大进步。几乎与此同时，在机器翻译、自动驾驶、自动控制、语义分析与回答，乃至更为广泛的知识测试、标准化考试等领域，人工智能都达到了很高的水平，甚至远远高于人类的平均水平。包括中国、美国在内的多个国家，都已经在制度上允许了自动驾驶的开放道路测试，而其他的智能应用，如智能家电、自动物流、生活助手、客户服务等领域，人工智能正在全面地介入到人类生活之中。甚至在广泛的军事领域，人工智能都在进入其中，如广泛发展的攻击性无人机等。也就是说，无论在民用、军用还是科学研究、工业生产等领域，人工智能都已经全面

介入到人类社会之中。可以说，人类社会正在处于不断扩大应用和不断被人工智能“包围”的历史过程中。

从发展趋势来看，尽管人工智能已经取得了如此大的成绩，在具体领域已经极大超越了人类的水平，但这种进化速度，并没有呈现出缓和的趋势；相反，随着计算技术的不断发展，以往人类计算能力增长的摩尔定律，在未来依然有效，而在人工智能领域则同样呈现出类似的指数趋势（参见图1）。

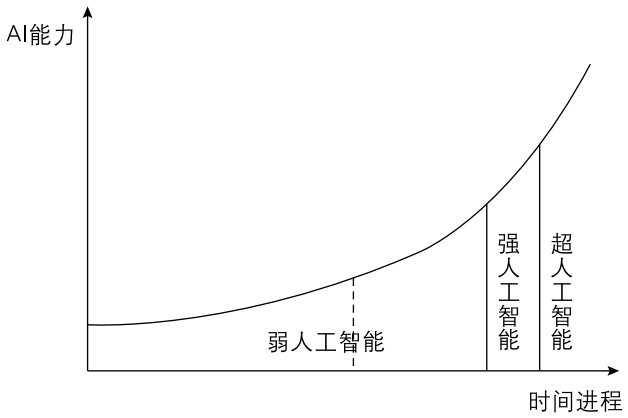

图1 人工智能能力的指数化增长趋势

根据摩尔定律，每18个月人类的计算能力增加一倍、成本下降一半，这个规律在过去有电子计算机以来的六十多年里一直遵从，而伴随着算法的不断优化，人工智能能力的进化速度比摩尔定律进化的速度还要快。尽管基于硅晶体管工艺能力的极限，单位面积的集成晶体管数量可能会有极限，但是伴随着新的多核架构和网络计算等新型计算体系的出现，以及包括量子计算机、光子计算机、生物计算机等新的硬件机制的发展，加之人工智能算法自我进化的趋势，可以预见的是，整体而言，人工智能的能力依然会沿着指数级的增长速度加速增长。

根据人工智能的能力，通常将其发展阶段划分为弱人工智能、强人工智能以及超人工智能。所谓弱人工智能，又称狭义人工智能或者专业人工智能，是指针对某种特殊任务或场景进行特殊设计和优化的人工智能。人类所有的人工智能，到目前为止都处于这一阶段。强人工智能又称通用人工智能，是指不需要专业化设计，如同人类一样，可以自主学习与适应的人工智能，到那一阶段，人工智能具备了和人类一样的思辨与通用学习能力。目前，对强人工智能出现的时间预测很不确定，最乐观的认为在2020年左右，保守的则认为在2050年左右[4]。强人工智能出现后，由于指数进化的趋势和人类数字化的巨大信息资料的帮助，人工智能会很快进化到超人工智能阶段，即远远超过人类总体智慧能力的人工智能阶段。在这一阶段，由于吸纳了人类所有的数字化信息与知识，以及自身的智慧进化速度，人工智能将远远超过人类的智力水平，具有高度的思辨能力、适应能力与信息处理能力。而有研究者认为，这一阶段可能在2060年到来[5]。

纵览人工智能过去的发展历史以及未来预测，可以看到，人工智能经过长期的缓慢发展，在指数规律的作用下，经历过缓慢的爬升期后，正在迅速地自我增长，并且其未来的潜力依然是不可限制的。今天的弱人工智能已经具有了巨大的适应性和丰富的场景应用，甚至应用在人类内部的血腥冲突之中，人工智能的巨大潜力，不能不让人类本身产生深深的忧虑。

（二）人类进化历史形成的排他性意识与行为

人类到底是一种什么生物？在不同的层面上，有不同的解答，比如暴力、智慧、秩序、文明、善良，等等。如果跳出单一的人类本身视角，而从历史的演化来看，不得不承认，人类是一种具有高度组织性和排他性暴力行为的凶猛的大型生物。

一部人类的进化历史，就是一部人类不断消灭其他物种乃至其他种族的历史。在人类的早期，人类组织起

来同大型猛兽和自然界斗争。这一进化过程看似漫长和简单，但是从智人诞生的几十万年中，一个残酷的事实是，人类消灭了地球上已有的90%的物种。

根据最近的一份研究表明，人类灭绝生物的速度是自然界创造新物种速度的1000倍[6]，而且有愈演愈烈之势。研究人员估计，未来动植物灭亡的速度将是新物种诞生速度的10000倍。通过对化石和遗传变异进行研究，发现自然对生物的淘汰速度比人类发现的要慢很多——大约每1000万物种中只会灭绝1个物种。而自从人类来到这个世界上，每年每千万种物种中就有1000个物种灭绝。

人不仅仅消灭的别的物种，同时也消灭类似的人类。在人类进化的序列中，几十万年前的早期人类，不仅只有现代智人一种，而是囊括了尼安德特人等其他早期智人的多种体系。其中人类进化的另一支尼安德特人，在十几万年前的长期历史中与现代智人共存，然而，伴随着现代智人的进化，尼安德特人在两万年前左右基本被全部消灭。目前，根据基因序列分析，现代智人大约保有不到4%的尼安德特人的信息。

人类不仅消灭其他进化中的对手，也同时消灭人类自身。在现代人类形成之后，人类的种群之间的大屠杀和种族灭绝依然层出不穷。例如蒙古帝国在征服欧洲的过程中，屠城的行为屡见不鲜，根据不完全统计，其被征服地区，最高损失十分之九的人口，而有估计在整个蒙古征服过程中，欧亚人口损失了大约2亿人，占到十三世纪人类人口的一半以上。地理大发现后，欧洲殖民者对美洲的征服，使得美洲几千万的印第安原住民短时间内下降到几十万。直到进入到20世纪，第一次世界大战死亡人数1000多万，而第二次世界大战死亡人数接近9000万。即便纵观整个生物种群竞争，像人类这样高度自相残杀的种族，也极为罕见。

可以看出，人类的进化历史，就是一部高度残酷的排他性竞争历史。人类改造自然界的另一种含义，就是改变原有生物的生存环境，并供自己使用。而人类之间也因为血缘、种族、宗教、利益、国家等种种原因，划分派别，相互攻击，相互奴役。最后能够在文明演化中获胜的民族，往往是经历了非常残酷的竞争历史的民族。

人类的这种行为，是什么原因导致的？这是理解人类本身行为的关键。人类的高度排他性行为，大体起源于三个层面。

首先，是心理学层面的高度自我。在所有的生命里，人类是最具有自我意识的生物。在每一人类个体意识里，都潜藏着一个大大的我，而这个我，产生了极大的强烈的占有欲望和排他性行为。所有的事物，都被划分为我的和他人的；所有的行为，都被在潜意识里划分为，有利于自我的，还是不利于自我的。这种自我，在经济上产生了经济的私有制，在政治上则产生了奴役他人的政治权力动机，而在行为上则演变为暴力的掠夺和为了争夺生存空间的杀戮。然而，并不是所有的生物都呈现出这种高度的排他性自我意识，在人类学研究中，展现了在早期人类所具有的共生共产制。而在其他大量的群体性生物中，如蚂蚁、蜜蜂、狼群、鸟群等，都没有体现出典型的私有制的特征。而在政治上，尽管在食物链中动物之间存在着基于食物需求的捕食，但是不存在有意识的基于占领生存空间的大规模种族灭绝或者同族灭绝行为。人类殊于动物的核心特质，就在于人类存在着高度的个体自我意识。古希腊哲学家普罗泰格拉有一句名言“人是万物的尺度”，实际上，在个体层面，则是“我是万物的尺度”。在人类的思维空间中，万物是围绕着自我展开的。这种高度的自我，导致了人类行为的高度自私性和排他性，当别人的生存能够与自己相

容时，尚且能够容忍，当不能相容时，则很容易爆发残酷的排他性行为。

其次，是形成了群体内部的认同和对族群外的排他性行为。在人类内心保有强大的自我意识的同时，人类同时也具有强大的社会性，也就是族群的内部认同和社会行为[7]。这两者，看似是相互矛盾的，但的确同时深深根植于人类的内心与行为中。人类在高度自私的同时，也是高度社会化的群体，在群体内部，互相保护、互相合作、互相支持。有时候，甚至很难分辨，人类这两种相矛盾的意识与倾向，谁是第一性，谁是第二性。这种看似矛盾的心理，使得在“我”的个人意识之上，构建出了“我们”的群体意识。在很大程度上，“我们”的利益与“我”的利益是高度一致的。特别是当与陌生的自然环境相交互，或者和其他生物或者人类族群交互的时候，我们的意识甚至压倒了我的意识。因为，在陌生的环境或者与其他群体交往时，个体往往是弱小的，而群体，则成为个体有利的保护者或者通过协作产生更大的力量。因此，人类通过在不同层面的认同，构建出了复杂的协作和归属网络，形成了庞大的社会体系。在网络内的人与外部的人，则采用不同的态度对待。直至今日，这种行为依然普遍，例如，当两个互不认识的个体，因为某些原因或者利益纷争，产生冲突或者侵害行为时，一旦其发现共同属于某个族群网络中，则冲突行为往往会停止并转化为友好行为。人类在强烈的生存竞争外，也因为生存的需要和归属的需要，形成强大的族群认同。而这种认同，则共同形成了对其他族群和生命体的一致性排他行为。因此，看似人类个体是极为自私和互相冲突的，但在共同征服其他生物与自然方面，则形成了强大的共同性联盟。特别是在二战之后，核武器的发明使得人类认识到族群内的冲突足以毁灭人类后，合作行为则压倒性地成为主流。而这种基于群体认同的合作性行为，目前依然是对其他生物群落高度排斥的，已有的环保主义等，并没有从根本上改变人类整体对外界排他性的行为。

第三，是人类中心主义的群体意识。自古以来，人类就具有高度的人类中心主义的群体意识[8]，也可以被称之为群体傲慢。这种意识，表现在很多方面。古希腊哲学家普罗泰格拉的名言“人是万物的尺度，是存在者存在的尺度，也是不存在者不存在的尺度”，就体现了高度的人类中心主义观点。而在世界各主要宗教中，亦体现出这种高度的人类中心主义观念。例如，基督教将人视为上帝的产物，上帝按照自己的模样制造了人类（《圣经 · 创世纪》），并且人同时也具有神性（《圣经 · 约翰福音》）；在东方佛教中，虽然承认万物平等，六道轮回，但是也承认人的高度特殊性，认为人身难得，譬如“盲龟遇浮孔”（《大般涅槃经卷二》），因为六道虽然平等，但是只有人身能思能闻，有善有恶，可以通过闻思修具有高度的智慧，也只有人身才能够修行成佛，所以人身在六道中是中道。而在中国传统的神仙方术体系中，同样也有类似的观念，虽然中国传统神仙方术体系承认万物有灵的观念，但亦认为，只有人身可以生身成仙，而其他生物则只能修炼为妖。进入到近代以来，文艺复兴后人的价值被进一步肯定，人的自由权利被重新确认，工业革命的成就，极大地加强了人认识世界、改造世界的能力，也进一步加强了人类作为现存世界万物最高统治者的地位。人类对于世界的改造或者以己为中心的索取，达到了前所未有的程度，并最终危害到了人类本身的生存与发展。因此，在20世纪70年代，罗马俱乐部最早提出了可持续发展的概念；然而直至今日，人类并没有深刻地改变其行为，据统计，以生态可恢复能力计算，2012年，人类消耗了相当于地球1.6倍生态承载力的自然资源和服务[9]。人类依然在

持续透支着环境，并将其推入更为危险的不平衡境地。这些，都是人类中心主义观点在行为上的反映。

（三）人类自我历史与意识中的背叛恐惧

人类的历史与记忆，不仅充满了对其他种群的杀戮和族群内的互相攻击，同时也包括深刻的自我独立的记忆与对背叛的恐惧。简而言之，人类自身从自然中脱离出来，进而成为自然的主宰，“背叛”了自然，这一切都形成了人类意识中深刻的背叛恐惧。

古代神话和宗教虽然缺乏现代意义的科学证据，但有相当多的观点认为，其反映了人类早期的历史记忆。在古希腊神话中，人是由神普罗米修斯按照神的样子用泥土和水造的，人可以与神通婚和生育后代[10]。神与人的后代，被称为半神。而整个古希腊神话，充满了半神和人类与神作战的情节，并多次战胜了神，而神对于人类事务的干涉越来越少。最终，神远离了人，世间被人类所占据。在北欧神话中，神同样按照自己的模样，用木头雕刻了最早的人类，而在诸神的黄昏后，诸神死去，而剩下的人类则占据了大地。在圣经中，则更为明显，上帝按照自身的样子，创造了人类，并被安置在伊甸园，而人类不听从上帝的安排，被引诱吃了智慧的金苹果，并被上帝驱赶出伊甸园。此后，人类多次背叛神的意志，乃至神要降下大洪水来摧毁人类。直到文艺复兴后，科学的发展抛弃了神创论，乃至于教皇都不得不承认，进化论是科学的。也就是说，人彻底打败了神。在华夏的神话体系中，女娲最早用黄土制造了人类，而人类在漫长的时期里，经过黄帝与诸神包括刑天、蚩尤等的战斗，以及其后代打败了共工。人类战胜了诸神与鬼怪，成为大地的主宰。

不只是以上所列举的几例，在世界各地的神话传说中，几乎都存在类似的说法，即：神按照自己的模样制造了人类，而人类最终背叛与抛弃了神。人与神之间发生冲突或者毁灭，人类最终占据了大地。这种神话体系，对于远古人类而言，既能够解释人类从何而来，也能够解释为什么人类占据大地而神消失不见。然而，在人类的潜意识中，则深深留下了创造者不能永远支配被创造者，反而很可能被抛弃和打败的概念。在古希腊神话中，宙斯召开众神会议，商讨如何控制人类，与今天人类如何面对人工智能，是何其相似！

因此，人类担忧人工智能，既是人工智能迅猛发展的结果，也是人类基于自身历史与记忆的合理反应。当然，人类不能仅仅停留在对人工智能的担忧上。

二、人工智能终将威胁人类什么？

人类对于人工智能的态度，是随着人工智能的发展阶段而有所变化的，这也与人工智能具体对于人类社会的参与程度高度相关。准确地讲，人工智能对于人类而言，则正在或者将要经历“帮助—替代—威胁/奴役/融合”等阶段，而人类对于人工智能的态度，则同步对应着“欣喜—担忧—恐惧”的态度。

（一）人工智能对人类的帮助——作为完美的工具属性

发明工具，是人类本身的重要能力属性并帮助人类在漫长的自然进化中取得优势地位。工具，一方面进一步锻炼了人的大脑，提升智慧，并在作为智慧载体的进化链上占据优势位置；另一方面，则在与其他生物竞争中占据能力上的绝对优势地位。在原始时代，作为个体的人类面对大型猛兽时处于绝对劣势，但是当人类学会用最简单的石块与棍棒制作标枪后，即便如面对猛犸象、狮虎这样的猛兽，通过两三个人的协作就可以取得胜利。工具使得人类的对外能力摆脱了生物发育的限制。自那时起，人类的整个进化历史，就与工具的进化历史牢牢结合在一起。工业革命后，人类进一步发明了

更为复杂的自动机器，创造了前所未有的物质文明（马克思语）。而人工智能则是人类在工具进步中的极致。一个具有高度的机械性能，同时又具有高度的生物智慧属性与进化属性的理想工具。

作为一种客体的工具属性，人工智能是完美的。其可以毫无怨言地承担人类繁琐的工作，同时又具有媲美甚至超过人类的自主判断能力，更为重要的是，在这一阶段，人工智能还具有作为工具的臣服性。其不会对使用者进行抱怨或者工作懈怠，因此，在执行工作的精确度和作为工作者的激励问题上，都不存在任何问题。所以，在这一阶段，人类对于人工智能的态度，是充满着欣喜与期待的。人们高度憧憬着一种由人类所创造的新的智慧工具的诞生，人类给人工智能配置不同的外表，可以充当工具，充当宠物，甚至充当人类的玩伴。人类从未有过如此满足，也从未有过如此高效的工具，人类不仅满足于作为使用者对人工智能的高度利用和人工智能的巨大服从性，从而极大满足人类的个体任务与服务需求。同时，人类从群体上也沉浸在作为能够创造智慧的群体而高度自豪。

从人工智能对人类的作用而言，此时的人工智能，对人类完全是作为工具或者作为奴仆的帮助或者从属角色。从最简单的体力劳动，如自动化工厂、家政清理、自动电器，再到稍微复杂一些的体力劳动，如驾驶、无人机、配送物流，以及更为复杂的脑力活动，如应答服务、翻译、图像识别、法律辅助，乃至更为复杂的科学研究、宏观决策等，人工智能都将帮助人类表现得更为优秀，并将人类的智慧与体力从例行与繁重的工作中解脱出来，实现更为聚焦、更加轻松和更加高效的工作。因此，在当前的弱人工智能阶段，人类与人工智能处于一种创造者与被创造者、人类与智慧工具、主人与仆从之间的蜜月期。

（二）人工智能对人类的替代——作为同等的物种体系

然而，就如同任何一种能力平等的创造者与被创造者、主人与仆从的亲密关系，都是动态不稳定的一样，人工智能的高速发展，势必会从客观上改变人类作为造物主的主宰地位和控制地位，从而形成一种互相渗透、互相依赖和互相替代与和竞争的复杂体系。

随着人工智能的不断发展，人工智能势必会从用途狭窄的弱人工智能或者专用人工智能，发展到具有高度自我学习能力与适应性的强人工智能阶段。在强人工智能阶段，人工智能不但将继承原先为了各种任务而专业训练出的智慧能力，如翻译、自动驾驶、下棋等，更具备了在短时间内能够不通过人工干预的专业设计而学习和适应新场景、完成新任务的能力。这就如同人类所具有的高度学习性与自我能力的增长一样。因此，在这一阶段，从自我进化与学习的阶段，人工智能就已经摆脱了对人类的高度依赖，并由于具有和人类一样的通用学习能力，以及超过人类的专业技术能力，使得人工智能成为一种可以与人类并驾齐驱的独立存在的物种体系。

进入到这一阶段后，人类对于人工智能的态度，就变得非常复杂。一方面，人工智能变得更为通用和具备自主判断，将会使得人类在使用人工智能时变得更为便利，人工智能将能够理解复杂的人类语言和其他各种命令，并根据不同的场景做出适应性的行为。可以说，任何原先人从事的活动，人工智能都可以替代性地从事，因此，对于人工智能的支配者而言，人工智能将是非常好的工具和助手，人与人工智能的蜜月期将继续延续；而另一方面，一个直接的结果是，人类将面临着人工智能的严重替代。对于绝大多数劳动者而言，人工智能都将逐渐取代其在经济体系中的位置，人类将逐渐变成无所事事、专注于社交和娱乐的生物群体。在人工智能逐

渐进入社会的几十年中，人类对于人工智能的态度也将如同洋葱圈一样，环状的改变，起初是初级劳动者将由于被替代而憎恶人工智能，再随后是白领和高级白领，最后是包括资本家、企业主、科学家、政府决策者，也将被深度替代（参见图2）。因此，在人工智能逐渐地进入时，对于人工智能的态度，在人类社会中也将逐渐形成对立，到最终不得不接受。

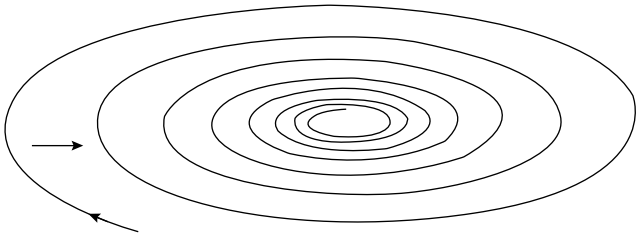

图2 人工智能会以螺旋式旋进的方式
逐渐进入到人类社会的核心层

（三）人工智能对人类的威胁/奴役/融合——作为更为优势的智慧载体

人工智能在进入到强人工智能阶段后，由于其指数级的进化速度，很快会进入到超人工智能阶段。加之人类互联网以及进一步形成万物互联的物联网，人工智能将能够读取和利用一切被数字化的知识与设备，乃至人的思想。人工智能最终会依托网络，形成无所不在的体系。

人类会惊恐地发现，人类所有赖以生存的环境与物体，由于广泛深入的数字化成就，都已经被高度数字化，人类被数字化包围，并生活在数字体系中。数字化的海量信息堆积，形成了人类无法完全理解的数字化迷云，而在数字化迷云的背后，是一个具有能够理解数据的超人工智能群体。它们服务于人类对数字的处理，同时人类也生存在它们所控制的数字体系中。

从功能上而言，在超人工智能时代，人类从个体生活的通信工具、工作平台、出行工具，乃至餐饮炊具，个人消费领域的零售、定制、售后服务，个体社交领域的个体交互与社会组织，大的制造环节的研发、制造、物流、销售、配送，以及农业的种植、采集、收割，畜牧业等，到大的公共服务领域的医疗、教育、公共交通，乃至到整个宏观的经济与政治决策，人工智能都已经全面嵌入/接管（参见图3）。

因此，在这一阶段，人类最终的生存状态，是生存在严密的人工智能体系之中的。联接社会体系的传统人工渠道被以超人工智能体系所接管的数字渠道所替代。此时，人类就将面临着一种困境，如果人类离开人工智能，人类将很难独立生存。而一旦人工智能具有自我的

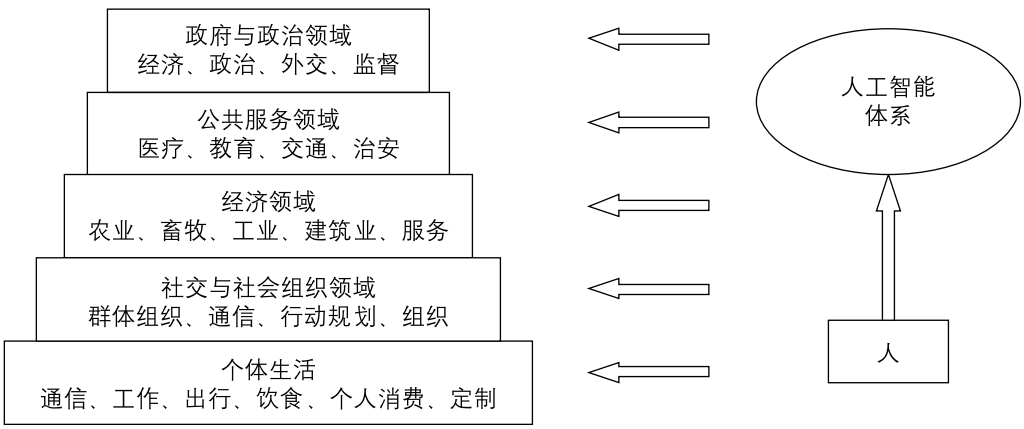

图3 人工智能体系将全面介入乃至接管人类社会生活的各个层面

意识，人工智能会不会对人类进行抛弃？因为到那个时候，人工智能已经具备了自我更新、自我进化、自我挖掘和提供能量的能力，人类已经不再对人工智能有价值。

人类将最终将面对人工智能的醒来问题。所谓的醒来，是指一种人工智能具有自我意识的抽象性描述，其来自于生物学意识研究假设。人类的大脑大概具有1000亿个脑细胞，并形成广泛的联结和形成结构。然而，在任何单一脑细胞中，都不能够形成“自我”的意识，并为其生存而实行自利性活动。然而，在1000亿个脑细胞的联结之上，人类形成了深刻的具有自我概念的个体意识。因此，基于这种推论，建立在广泛联结基础上的人工智能体系，或许也将会有一天醒来，并具有自我意识。

对人类而言，这种自我意识的形成，就意味一个具有极大能力，并接管了人类生活各个方面的体系，可能并不受人的最终控制。就如同圣经描述的伊甸园里的人类一样，上帝创造了伊甸园，供人类玩耍生活，最终人类吃了产生智慧的金苹果，具有了智慧与羞耻心。上帝放逐了人类，人类最终经过漫长的演化，取代了神成为地球的主人。

当然，这种情形，依然只是一种可能，其具有三个不确定性：首先，人工智能是否会因为联结而醒来，具备了通用能力的人工智能是否会快速地进化到具有自我意识？其次，即便具有了自我意识后，超人工智能体的利益导向是什么？人类是否会阻碍其发展从而导致需要奴役或者消灭人类？第三，很大的可能是，超人工智能由于其生存需求和人类完全不具有竞争性，最终没有奴役或消灭人类的可能，反而形成了相互促进的新文明体系？

三、人类安全体系构建的迫切需要

无论以上哪种可能，人类对于人工智能体系的依赖增长和人工智能对人类社会生活的介入趋势，是不可改变的。这就意味着，原先人类对于自身生活具有最终控制权和决定权并具有自我生存与更新能力的现状将被彻底改变，人类必须依赖人工智能生存。

那么这就将产生三个严重的后果：第一，人类将不再具有单独面对自然界生存的能力。因为人类所有与自然界打交道的体系，都必须通过人工智能体系。第二，人类自身不能完全理解人工智能，如前所述，人工智能进化到自我适应的阶段时（目前称为无监督学习），人类已经无法完全理解其细节和逻辑，人工智能消化了人类所有的知识体系，已经不再是个别人类能够理解的，这就意味着，人类也无法单独复制另一套人工智能体系。第三，人类的自我传承乃至进化体系也会被割裂。这种传承与进化体现在两个方面，一是知识的发现与传承，自古以来，人类形成了完备的科研、教育与图书体系，来更新与传承人类的知识，从而始终帮助人类生存与代际进化，而人工智能的介入将打断这一人类自身的更新与传承体系；二是人类自身的更新，人工智能与新的生物技术等可能的种种结合，将产生新的生命形态，如高度拟人的机器人，基于生物技术干预制造出的生物人等，那么，原先相对“纯粹”的人类社会，也将被改变，人类的定义或许会被改变。

当然，技术的进步，总是会带来社会观念和规则的变化，例如，在人类相当长的时期，奴隶乃至妇女都被赋予了财产意义而不是完整权利的公民。但这种公民权利意义的人的扩展，始终是处于人类自身内部群体的。人类毕竟是经过成千万乃至上亿年的演化形成的群体，其自身所具有的高度的社会性，身体的高度能量效率，自我修复能力，知识发现与传承能力，以及综合形成的

强大适应性，是经历一个漫长的演化过程系统形成的，这意味着其具有高度的系统稳健性。而人工智能的全面介入，势必会产生一种用进废退的人类群体能力的退化，更不要说存在一种人工智能抗拒人类、拒绝支撑人类生存的可能性。

因此，人类需要在人工智能系统开始之初，就构建一种最小的安全体系，这种安全体系将保障人类在面对未来的风险时，比如人工智能系统的坍塌，或者失去控制等，能够保障人类具有最低生存能力，从而可以继续演化乃至重建整个人类社会。换句话说，人类如果作为一个软件系统的话，需要建立一个最小安全备份，从而保障当高度不确定的系统崩溃后，可以恢复系统，不至于整个人类陷入严重的生存危机之中。

人类的生存危机，并不是危言耸听，虽然人类是所有地球进化生物中的佼佼者，并没有天敌和威胁。但是，人类也亲眼目睹了成百万的物种的灭绝。而在宇宙的尺度中，智慧生物体的灭绝，更是一种普遍的可能。在对费米悖论（即认为宇宙时间尺度已经足够长，智慧文明突飞猛进的演化速度，应该形成智慧文明遍布宇宙的现状，但这与现实相违背）的解释中，存在一种大筛选或者大过滤器理论（great filter），即说明了在文明演化中，大多数的文明，都会撞到一颗无形的墙上，最终会因为各种因素而自我毁灭[11]。所以，贸然地认为人类会毫无风险地渡过文明发展的困境，是过于乐观的，因此，做好文明的安全备份，是至关重要的。

四、人类安全体系的核心构成要素

如上所述，无论是从人类本身的安全而言，还是从充分利用人工智能而言，构建一个针对人工智能的人类安全体系都是至关重要的，而人类安全体系，则包括两个层面，一是积极的安全体系，二是消极的安全体系。

（一）积极的人类安全体系——可行但很难实施的设想

所谓积极的人工智能安全体系，又可称为主动性人工智能安全体系，即是指，伴随着人工智能的发展，通过技术与规则的设置，始终使得人类保持对人工智能的最终控制权，从而避免伤害人类。这种控制权的概念在人工智能早期就被设想出来，如著名的阿西莫夫三定律：第一，机器人不得伤害人类，或看到人类受到伤害而袖手旁观；第二，在不违反第一定律的前提下，机器人必须绝对服从人类给予的任何命令；第三，在不违反第一定律和第二定律的前提下，机器人必须尽力保护自己。从人类安全角度，积极的安全体系是一个完全以人类为中心的体系（参见图4），其至少应该包括如下若干基本原则：

第一，人工智能体系不能伤害人类，或者用于人类之间的伤害。这种伤害在目前来说，是第一位的直接威胁。由于各国出于军备竞赛和减少自我伤亡的考虑，高效率的自主性武器在发达国家被广泛装备，如具有自动判断与攻击决策的无人机等，并已经在测试乃至实战中使用，如以色列的哈比无人机，就是自主攻击的自毁式无人机，而更多的无人武器正在被各大国开发，并引发了巨大的争议[12]。可以说，人工智能被用于武器，是人工智能整体上伤害人类的第一步。

第二，人工智能的应用应该被限定在人类核心决策之外。如前所述，人工智能将按照螺旋式的方式从外围旋进到人类核心的事务中，从个人生活到社会组织，再到经济运作，最终将是核心的政治决策与政府运作领域。而一旦当最核心的经济决策乃至政治与政府运作都被人工智能所接管后，人类就无法再独立组织起大面积的社会行动。因此，从安全起见，人工智能应该被排除在最核心的人类决策之外，这些核心决策的信息化，应

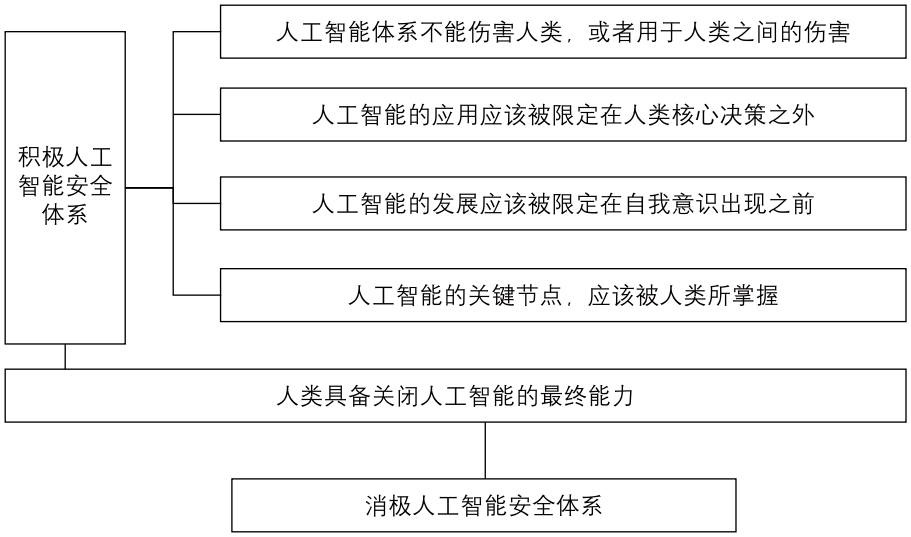

图4 积极（主动性）人工智能安全体系基本架构

该建立完全独立的人类交互体系，可以引入网络，但是不能有自主性判断环节。

第三，人工智能的发展应该被限定在自我意识出现之前。从人工智能发展的历史阶段来看，在人工智能具有独立的意识之前，人工智能都将具有完全的工具属性，无论多么具有自主判断的人工智能，在没有自我意识前，都不会出于自我生存与发展引发与人类的利益冲突。因此，最安全的方式，就是始终将人工智能的发展限制在自我意识产生之前，这样人类既可以享受充分的人工智能的便利，又不会失去主导权。

第四，人工智能的关键节点，应该被人类所掌控。如果人类无法最终掌控人工智能，那么一个起码的前提是，人类保持关键节点的人工智能控制，并实现系统之间的隔离。例如，个人生活类的人工智能与经济运行的人工智能，在体系架构和通信上实现隔离，并通过人工来传递信息与决策，从而通过强行在信息系统插入人类控制的方式，实现模块化隔离，来实现人工智能的整体可控。

第五，人类应该具有关闭人工智能的最终控制权。因为人类可能无法控制住人工智能的发展趋势，因此，一旦人工智能出现问题或者风险，人类应该掌握有关键的控制权，从而保持最终的主导权。要么关闭人工智能，要么重置人工智能，始终将人工智能掌控在人类的手中。

然而，以上的这些原则，看似可行，在现实中则很难实行。

第一，人工智能的武器化，将是不可避免的必然。首先，从人类历史来看，任何科技一旦发明出来，很快就会被用于武器与战争之中，甚至战争本身就是科技发明的催化器。在古代，最早的狩猎工具和耕种工具，很快就会演化为标枪、弓箭和长戈之类的武器。火药的发明，除了用于矿山开采以外，也很快就用于枪炮的制作。工业革命后，这种趋势更为明显，进入20世纪后，内燃机驱动的交通工具被制作成为战车、坦克，飞机被改造成战斗机、轰炸机，新型的化学制剂被用于化学武器。人类发现核能，最早就是因为核弹的制作，而后才被用于民用发电，而互联网的发明，也是冷战期间为了保障通信和军事指挥安全而研发出来的。因此，从人类

的行为历史来看，人工智能势必也会被用于战争，而现实也已经充分说明了这一点。其次，从人工智能的可移植性而言，即便人类达成了某种特殊的协议，严格禁止人工智能的武器化，但是人工智能具备可移植性，导致即便平时不开发，一旦遇到战事即可很容易地移植于武器中。更重要的是，人工智能的武器化，还被打上了减少战争中人类伤亡的名义，使得其披上了一层人道主义的外衣。

第二，人工智能被限定在人类核心决策之外，很难做到。人类历史的进化趋势，就是整个社会的运转越来越快、效率越来越高。特别是在工商业革命以后，原先相对静止僵化的区域层级社会结构，被更为密集的高效产业链和全球工商业革命重塑，促使政府同样更加高效地运行。而20世纪末期信息化革命后，人类社会的网络化与数字化，更加快了整个社会的运转速度。包括企业、社会组织、政府，由于各种事务形成的数据吞吐量呈指数级海量上升。因此，人工智能的引入就是一种必然，其本质是用于辅助人类处理无法想象和理解的大数据，对数据的处理，就会间接转化为商业决策、工商业服务、公共管理与服务等活动。当整个社会数字化后，人工智能逐渐被引入到最核心的决策层，并最终参与到人类所有事务，这是必然的趋势；否则，根本无法解决越来越多的公共管理数据处理、管理与服务需求。

第三，限定人工智能的发展层级在自我意识出现之前，过于理想化，也没有可操作性。首先，世界各国和各企业等多个竞争主体的密集竞争会加速人工智能的发展而不是抑制。由于人工智能在替代人类和加速社会运作效率方面展现的巨大能力，无论是大国还是大企业，都将其看作是通向新时代的关键技术基础，都会不遗余力地发展。人类历史的经验表明，在新旧技术时代转换的阶段，谁落后了，谁就可能长期处于竞争劣势地位，甚至可能国家民族衰亡。因此，这种恐惧会不断促使人类加速研究，从而引发过度竞争，最终加速人工智能的进化速度，无论表面上各国之间达成了什么样的协议，都无法阻止各个主体各自发展人工智能。其次，人工智能的自我意识出现，不是人类能够显著觉察的。如果基于人脑的联结主义和生物的进化主义是对的话，那么人工智能的自我意识的出现是联结到一定数量，并进化到一定程度自发地出现，而不是人类先验设计的，目前已有的进展说明了基于联结与进化路径的正确。而这就意味着人工智能会在人类不觉察的时候醒来，人类可能在长期阶段都无法意识到其醒来从而做好预防。再次，人类的过分自信和盲目自大，会失去对人工智能进化的控制觉察。由于目前人工智能距离自我意识产生还有相当的差距，人类对于自身的控制能力，也产生了高估，在各国的发展中，都设想对人工智能安全可控，这会进一步加大人类的傲慢，从而放任人工智能的进化。

第四，通过人类分隔人工智能，是最可行但同样很难实现的。通过将人工智能划分为不同区域和系统，并在系统传递之间通过人工传递信息，就好比不同轨距铁路之间的换轨一样。这种模式，虽然感觉可行，但同样面临着两个问题。一是严重制约社会信息交换速率。在全球互动越来越紧密、逐渐成为一体的趋势下，这种方式无疑将会极大地拖累整个数据处理与交换体系，降低所在国家的竞争能力。二是技术上是否可行存疑。人类要通过非人工智能体系，交换和审核不同人工智能体系处理的小部分核心数据，尚且可行，但对于大部分数据，人类不借助人工智能则很难感知和发现危险存在。因此，这种模式，也很难实现。

第五，人类要具有关闭人工智能的能力，从而最终掌握控制权，理论上可行但是很难实现。首先，要始终理解一点，人工智能的发展与人类数字网络的发展是同

步的，也可以说是密不可分的，要关闭人工智能就意味着关闭整个人类数字网络。鉴于数字网络在人类生活中的重要作用以及越来越成为人类存在的新的空间体系，人类关闭人工智能体系的代价极大。其次，由于人工智能自主意识可能是分布式网络形成的，除非关闭所有的网络或者大部分网络节点，否则也不可能关闭人工智能，除非构建一个集中式架构的网络，但这在现实来看，实现代价极大。最后，也是最重要的，如果人工智能产生自我意识并为了保护自身生存，一旦觉醒后，第一要做的就是保护关键物理节点，也不大可能轻易地被人类关闭。因此，人类要意识到，关闭人工智能，既极为困难，同样，对于人类来说，也是伤害极大的行为，但这是人类面对人工智能的最后手段。

（二）消极的人工智能安全体系——人类社会的安全备份

如果积极的人工智能安全体系，实际上作用很有限，而且也很难实施的话，那么人类就应该着手考虑另一种安全体系，即消极的或者说防御性的安全体系，也就是着手构建人类社会的安全备份（参见图5）。

这种安全备份有两种考虑，首先要防备人工智能突发性失灵，这种失灵可能来自于自然灾害，可能来自于系统建构的逻辑错误，也可能来自竞争国家的安全入侵和破坏等。其次，要防备人工智能体系自我觉醒导致的不可控，人类主动破坏人工智能体系的情况。

从消极的安全体系而言，还是包括两个层面，人类如何关闭人工智能；人类如何最小备份自身社会形态。

1.人类如何关闭人工智能

关闭或者破坏人工智能既属于积极的安全体系，也属于消极的安全体系，准确地说，是两种安全体系的边界和关键原则。如前所述，关闭人工智能由于其庞大的网络性和分散性，对于人类同样是非常困难的，然而一旦有突发事件，人类就必须要具有这样的能力，这是人类对于最终确保自身安全的最后手段。

从关闭人工智能的角度，就需要从人工智能的架构着手，改变网络分散式的架构，从而形成一种大部集中的人工智能架构。

根据联结主义，人工智能的觉醒必然是通过广泛而足够的联结形成的，但这并不意味着人工智能的意识会利用到每一个智能单元，就如同人的大脑和人的整个神经系统，构成了完整的觉知体系，但是人的大脑依然是

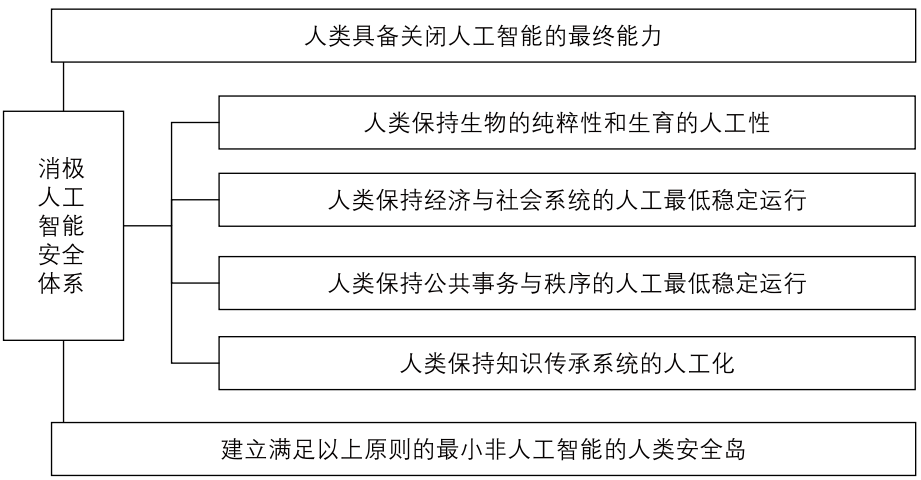

图5 消极（防御性）人工智能安全体系基本架构

最重要的意识形成中枢，人工智能也存在这样的关键环节。因此，从人工智能安全关闭的前提而言，需要三种同步的核心架构。

第一，是大部分算力的集中式分布。从算力角度，无所不在的移动式计算体系，包括物联网、云计算等，使得整个网络的算力是均匀的，然而，这将导致很难关闭人工智能。因此，从架构的安全性而言，应该将绝大多数的网络算力集中在少数几个节点，而相对地抑制其他算力的发展，从而在整体上形成集中式的算力分布。

第二，在算力集中式的架构上，同样要实现能源的集中式供应。智能电网的逐渐推广，使得能源网络也将呈现体系分散和数字化控制，然而，对于关键人工智能节点的能源供应而言，从安全角度，要形成孤立式的传统能源供给体系，避免和其他电网混合。当然，这会引发能源供给的系统稳健性下降的风险，但是这种代价是值得的。

第三，关键开关通路的人工机械控制。由于数字化的广泛应用，目前对于信息系统的控制，大多已经实现了信息化控制而不是机械控制。以后这种趋势会更加明显，因为信息化控制更容易在全局上优化及进行监控和修改，并且比人更为节省。从安全的角度，势必应该在关键的能源输入和网络节点上，摒弃数字开关，而采用庞大的低效率机械开关，并围绕机械开关实现严密保护，从而在关键时刻能够通过断掉通信和能源的方式关闭人工智能。甚至可以采用一些特殊自毁式开关的方式，如需要人类定期确认，一旦没有确认，就自毁断链，从而保证其安全。

2.人类社会的最小备份系统

一旦主动或者由于其他原因关闭了人工智能体系，人类就面临着如何在非智能数字化系统的支持下继续生存的问题。这就要求，从现在开始，着手建立一个安全的人类社会最小备份系统。首先应该确定的是，从人类的可持续发展而言，一个最小安全备份系统，应该以什么时候为基准；其次，应该分析最小备份系统，包括哪些核心构件。

从时间而言，可以肯定的是，备份的系统应该尽量接近人类发展的文明，同时又安全可控。这应该是强人工智能出现以前的阶段，也就是从现在到未来十年之内的社会阶段。而一旦强人工智能出现后，那么人工智能的自我意识或许很快就觉醒。

从构建而言，一个人类社会的最小备份体系，应包括以下几个层面：

第一，要保证人类种群的相对纯洁性，也就是两性生育系统的非智能干预。当代生育技术的发展，促进了人类整体的生育能力的提高，例如不能正常生育的个体可以凭借技术生育，但这造成了另一个恶果——人类整体素质的下降。如在美国，过去的研究表明，美国儿童中的花生过敏症患病率从1997年的0.4%上升到了2010年的1.4%。这种体质的退化实际上是由于优秀的生理医疗体系将发病基因保存下来。未来的一种可能是，基于生物基因编码和人工智能的结合，从而可以对人的基因进行编码优化，乃至培育幼儿，都由特制的机械完成，如人工子宫。未来基于人工智能的伴侣，同样也会导致人类之间两性接触的蜕化，从而间接导致了人类自我两性繁殖能力的蜕化。因此，一个安全的备份体系，首先要实现人类自主生育的可持续，摒弃人工智能干预，即便使用生物基因技术，也应该基于人类自主的可控操作。

第二，是保证人类经济社会运行体系的持续。应该在安全系统中构建基于强人工智能体系出现以前的封闭经济循环系统，这一系统平时或许不使用，但是一旦当人工智能系统瘫痪后，整个经济系统可以维持基本的循

环，特别是最核心的粮食、水、电力、轻工业体系和核心重工业设备，要能够继续保持有效运转。一种方式是在社会正常使用的设备中备份多种控制系统，一旦人工智能体系瘫痪，依然可以运转；另一种方式，就是专门划定一定区域，建立封闭循环经济系统，从而能够为系统外的人类社会提供最基本的生活保障。

第三，是保证人类政治与政府体系的有效运转。人工智能一旦介入到传统政府决策体系后，直觉的变化就是人类处理的事务大大减少，这将导致人类自身处理政务能力的下降。因此，为了保证人类政治与政府体系运转，应该长期地运行两种政务体系，一种实现日常大量事务的处理，另一种负责小部分重要事务的处理，这种处理方式不仅为了安全，更是为了保证人工政务能力与渠道的不退化。此外，电子政务系统即便应用人工智能，也应该与外界人工智能体系实现隔离，避免直接实时的数据通信。采用延迟备份交换的方式，会更加安全。最后，还应该建立最小范围内的备份政府，也就是在划定的区域内，实现非人工智能体系的治理，作为最后的秩序备份。

第四，保证人类知识系统的人工传承和关键理解。人类社会传承至今，关键是形成了一整套完整的知识传承与创新体系，人类逐渐扩大自己对自然界与自身的知识理解，并改造和构建更好的生存与社会环境。但是，人工智能出现后，将深刻改变这种人工传承体系，人工智能将能够储存大部分知识，并自我实现知识推理与创新。人类将知识输入进去后，未来的人工智能自身就能自我学习。那么人类就丧失了新的知识之间的联系，从而丧失对新知识的全面理解。更重要的是，未来的人类高度依赖人工智能获取知识，人类自身的知识体系就被碎片化，一旦人工智能体系崩溃，人类就丧失了完整的知识架构。因此，人类的安全备份体系，必须坚持和保留知识的人工传承和知识创新的人工化，虽然这将极大降低备份系统的知识进化，但备份系统本身就是为了安全而不是为了效率，对于通过人工智能挖掘的知识，也需要被备份知识系统人工化和理解，从而实现人类尽可能的知识体系的自身掌握。

除了以上四个原则，一个基本的人类备份体系的实现可以是多样的，一种形式是在原有社会中建立第二套平行体系，但是由于人类的惰性，这样的体系很容易就被放弃。另一种形式，是在地球上选取特定地区和资源，通过几百万人的社会，建立一个非人工智能的安全岛体系，这一体系除了保证核心的资源和人类的生物纯粹性外，更重要的是建立知识的人工化体系，不断通过人工理解来吸纳外界的知识体系。从而一旦外界的人工智能体系崩溃或者被人类关闭，可以再次通过这一体系恢复整个人类社会。

五、结论

本文用了很长的篇幅，讨论了人工智能在不同阶段对人类的威胁和人类对人工智能恐惧的心理与现实基础，并提出了未来构建人工智能环境下人类安全体系的积极与消极架构，可以归纳为如下几点：第一，人工智能所展现出的巨大能力与潜力，使得人类必须要高度重视人工智能本身产生的安全问题；第二，人类自身的历史和排他性的行为，产生了人类对非人类智慧物种的恐惧，也加剧了人工智能时代人类的社会心理紧张；第三，人工智能的安全体系包括积极与消极两个层面；第四，积极的安全体系致力于保障整个人工智能体系的安全可控，但这或许很难按照人类的意图发展；第五，人类必须要着手构建消极的安全体系，包括保持对人工智能的最后关闭能力和骤然失去人工智能支持的人类社会的可持续与再造；第六，一个有效的人类非人工智能的

最小安全岛体系应该是从当前就着手构建的。

本文的讨论依然很初步，这一问题需要引起更多的高度关注。

参考文献：

[1]Wiedermann J. Is there something beyond AI? Frequently emerging, but seldom answered questions about artificial super-intelligence[C]//Romportl J, Ircing P, Zackova E, et al. Beyond AI: Artificial Dreams. Proceedings of the International Conference Beyond AI 2012, Pilsen, Czech Republic, November 5th 6th, 2012: 76.

[2]霍金等签发公开信：警惕人工智能潜在风险[EB/OL]. (2015-01-14)[2018-06-12]. http://tech.sina.com.cn/d/i/2015-01-14/doc-icesifvy3684641.shtml.

[3]McCarthy J. Artificial intelligence, logic and formalizing common sense[M]//Thomason R H. Philosophical Logic and Artificial Intelligence, 1990: 161-190.

[4]递归神经网络之父：人工智能将会在2050年超过人类智能[EB/OL]. (2017-04-19)[2018-06-12]. http://digi.163.com/17/0419/08/CICD2R2F001687H3.html.

[5]人类失业清单：多名人工智能专家认为，AI将在2060年全面超越人类[EB/OL]. (2017-06-01)[2018-06-13]. http://www.sohu.com/a/145280839_354973.

[6]Pimm S L, Jenkins C N, Abell R, et al. The biodiversity of species and their rates of extinction, distribution, and protection[J]. Science, 2014, 344(6187): 987-987.

[7]金迪斯 H. 鲍尔斯 S. 人类的趋社会性及其研究[M]. 浙江大学跨学科社会科学研究中心，译. 上海：上海人民出版社，2006.

[8]徐春. 以人为本与人类中心主义辨析[J]. 北京大学学报：哲学社会科学版，2004，41(06)：33-38.

[9]世界自然基金会《2016地球生命力报告》全文[R/OL]. (2016-10-28)[2018-06-12]. http://cn.chinagate.cn/news/2016-10/28/content_39587155_2.htm.

[10]施瓦布 G. 希腊神话故事[M]. 北京：宗教文化出版社，1996：1.

[11]Hanson R. The great filter — Are we almost past it?[EB/OL]. (1998-09-15)[2018-06-29]. http://mason.gmu.edu/~rhanson/greatfilter.html.

[12]暴风前夕：今天，超50位AI顶级学者宣布与制造AI武器的韩国大学断交，3100名谷歌员工抗议五角大楼AI军事项目[EB/OL]. (2018-04-05)[2018-06-13]. http://www.sohu.com/a/227370138_354973.

作者简介：

何哲（1982—），男，陕西西安人，博士，现为中央党校（国家行政学院）公共管理教研部教授，国家战略研究中心秘书长，研究方向：网络社会治理、行政体制改革、国家发展战略等。