Research into risks from artificial intelligence

Lijun Chen[1, a], Xiaoru Chen[2,b]

South China Institute of Software Engineering of Guang Zhou University, Guangzhou, 510990, China

South China Institute of Software Engineering of Guang Zhou University, Guangzhou, 510990, China

[a]email: 372158286@qq.com [b]email: 479170369@qq.com

Keywords: artificial intelligence; Advantages; Risk management; Artificial consciousness; outlook

Abstract: At the moment, artificial intelligence and increasingly sophisticated algorithms have influenced our lives and civilizations more than ever. The fields of artificial intelligence applications are diverse, potentially far-reaching, and because of recent improvements in computer hardware, some artificial intelligence algorithms have surpassed the capabilities of today's human experts. With the improvement of artificial intelligence, its application field will continue to grow. From a specific point of view, the relevant algorithms are likely to start optimizing to reach a higher degree, and, one day, artificial intelligence can reach the level of human intelligence. This technological advance may bring us unprecedented ethical challenges. Many experts believe that, at the same time as global opportunities, artificial intelligence has brought more global risks than nuclear technology (whose risks have been severely underestimated before development). In addition, the scientific risk analysis suggests that the high potential damage of artificial intelligence should be very serious, even if the probability of occurrence is very low.

基于风险的人工智能研究

陈立军[1, a]，陈孝如[2,b]

[1]广州大学华软软件学院软件工程学系，太平镇，从化，广州，中国

[2]广州大学华软软件学院软件工程学系，太平镇，从化，广州，中国

[a]emai:372158283@qq.com，[b]emai:479170369@qq.com

Mobile: 139022645336 author: lijun chen

关键词：人工智能；优点；风险管理；人工意识；展望

摘要：目前，人工智能和越来越复杂的算法，比以往任何时候都更影响我们的生活和文明。人工智能应用的领域是多样的，可能影响深远，而且由于最近计算机硬件的改进，某些人工智能算法已经超过了当今人类专家的能力。随着人工智能能力的提高，它的应用领域将继续增长。从具体的角度来看，相关的算法很可能会开始优化，以达到更高的程度，并且，总有一天，人工智能能达到超人类的智力水平。这种技术进步可能会给我们带来前所未有的伦理挑战。许多专家相信，在全球机遇的同时，人工智能带来的全球风险超过了核技术（其风险在发展之前被严重低估）。此外，科学的风险分析表明，人工智能的高潜在损害应该非常严重，即使它们发生的概率很低。

1. 人工智能介绍

知识的追求贯穿人类的历史，贯穿始终。每当社会在其动态和结构上发生重大变化时，这通常是新技术发明的结果。在 200 万年以前，第一次使用石器的历史时期，一部分聪明的人发明了艺术，并开始在洞穴墙壁上作画。又过了三万年，耕种农业和永久定居的兴起。第一个符号出现在那之后的几千年，紧随其后的是最初的书面文字。大约四百年前，发展开始加速。显微镜是在十七世纪发明的;19 世纪的工业化使 100 万人口的第一个城市得以实现;在上个世纪，原子被分离，人类登上月球，计算机被发明。从那时起，计算机的处理能力和能源效率每隔一段时间就增加一倍[1]。然而，尽管技术进步呈指数级增长，但人类智力的能力却并非如此。

近年来，无数著名的科学家和企业家都对人工智能的重要性提出了警告，而政策制定者应对人工智能研究提出的挑战非常重要。

在某些领域，人工智能已经在多个场合达到甚至超过了人类水平。1997 年，计算机深蓝击败了当时的世界冠军加里·卡斯帕罗夫[2];2011 年，沃森在基于语言的游戏《危险边缘》中击败了两名最好的人类球员。与此同时，人工神经网络可以与人类专家在癌症细胞的诊断上进行竞争，在识别手写汉字的过程中也或多或少接近人类水平[3]。早在 1994 年，一个自学的 backgammon 项目通过寻找前所未有的策略，达到了世界上最优秀的球员的水平。现在，甚至有一些算法能够独立地学习许多不同的游戏，从而达到(或超过)人类的水平。随着这些发展，我们正在慢慢地接近一般智力，至少在原则上可以独立解决各种各样的问题。

权力越大，责任越大。技术本身只是一种工具;重要的是我们如何使用它。使用现有的人工智能系统已经给我们带来了相当大的道德挑战，这将在本文的下一节阐明。接下来的章节将概述经济自动化的发展，并解释人工智能研究将导致劳动力市场的重大重组。最后两章将探讨人工智能研究与人工智能可能产生的长期和存在风险。下图是中国人工智能技发展现状。

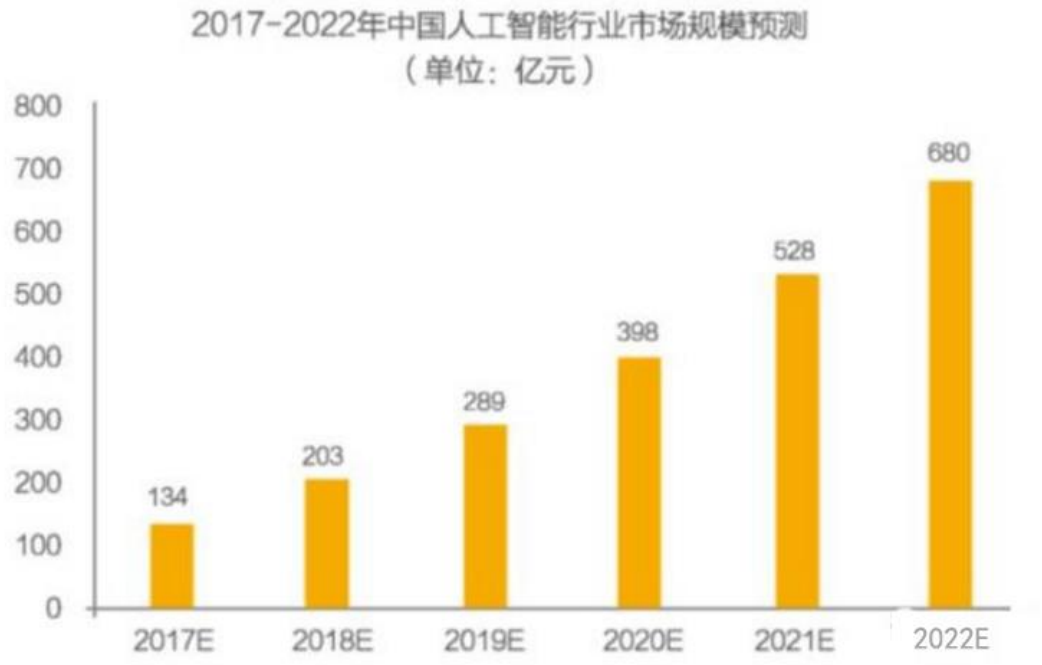

图 1 中国人工智能技发展现状

2.当前人工智能的优点和风险

我们的个人生活和整个文明都可能被算法和特定领域的人工智能所统治。下表是很有可能被人工智能取代的职业：

表 1　6 个风险性最高的工种

工作岗位	被取代的可能性
电话销售	99%
贷款专员	98%
收银员	97%
律师助理	94%
出租车司机	89%
快餐厨师	84%

众所周知的例子，包括智能手机，空中交通控制系统和互联网搜索引擎、金融市场也依赖于算法，这些算法对于任何一个人来说都过于庞大而复杂，无法完全理解。在大多数情况下，这种算法的运行没有发生任何意外，但总有可能发生“黑天鹅”事件[4]，可能会使整个系统陷入混乱。我们已经目睹了这样一个事件：2010 年，美国股市出现了一场意外的“闪电崩盘”，让金融世界目瞪口呆。由于计算机算法以一种无法预见的方式与金融市场互动，导致重要的股票损失了超过 90%的价值，然后很快又回到了最初的价值。如果这样的事件发生在军事背景下，类似的“回到初始条件”将造成不可预料的事故。为了防止这种破坏性的失败，通常建议将更多的资源投入到人工智能的安全性和可靠性上。不幸的是，目前的经济刺激似乎更有利于人工智能的能力，而非安全性。

2.1 人工智能建设的四个标准

安全对于任何机器的建造都是必不可少的。然而，当构建特定领域的人工智能能时，新的伦理挑战就出现了，—直到现在，这些工作都是由人类完成的。例如，一种判断银行客户信用评级的算法可能会做出对人群中某些群体的歧视(而没有明确的编程)。即使只是简单地替换现有行动的技术也可能会给机器伦理带来一些有趣的挑战：例如，无人驾驶汽车提出了一个问题，在即将发生的事故中，哪些标准应该是决定性的？如果车辆确保乘客的生存高于一切，或在不可避免的事故发生时，优先考虑尽可能低的伤亡人数？因此，人工智能理论家尤科沃斯基和哲学家尼克·博斯特罗姆都提出了四项原则，应该指导新人工智能的建设：

（1）人工智能的功能应该是可以理解的；
（2）它的行为基本上是可以预测的；
（3）使负责任的专家能够及时做出反应，并在可能出现的失败时否决控制；
（4）人工智能应不受操控，万一发生事故，应明确责任。

2.2　人工智能的优势（领域特定）

在原则上，算法和领域特有的人工智能带来了许多优势，他们已经对我们的生活产生了更大的影响，并期望在未来不断增加的速度中继续这样做，只要采取必要的预防措施。在这里，我们将讨论两个具有指导意义的例子。无人驾驶汽车不再是科幻小说[5]，在不久的将来，它们将在商业上可用。由自主人工智能算法驱动的谷歌无人驾驶汽车于 2011 年在美国首次试驾。除了工作或放松的时间，无人驾驶汽车的第二个优势在于其更高的安全性。2010 年，全球有 124 万人死于交通事故，几乎完全是由于人为错误。因此，每年都可以拯救无数的人类生命，因为无人驾驶汽车已经比人类驾驶的汽车安全得多。自然，很多人对无人驾驶汽车持怀疑态度，主要是因为他们低估了汽车的安全性，同时高估了自己的驾驶能力。为了说明这一点，一项研究得出结论，93%的美国司机认为他们的驾驶能力高于人工智能司机，这在统计上是不可能的，不切实际的乐观和控制错觉可能也会使人们低估自己在方向盘后面的风险

[6]。

另外，医生也高估了他们的能力，在最坏的情况下会导致致命的事故。仅在中国，每年就有 19 万人因医生开错药而死亡，有高达 81%的门诊患者的处方是错误的。在此背景下，IBM 的沃森是一个受欢迎的开发。2011 年，这个人工智能在智力竞赛节目《危险边缘》中击败了最优秀的人类选手，一举成名。不过，在智力竞赛节目中，沃森不仅仅比人类更好，自 2014 年以来，医院一直能够雇用沃森的计算能力，用于癌症诊断和其他复杂的模式识别任务。因为“沃森医生”可以快速收集和结合大量的信息，它已经部分地超过了它的人类同事的诊断技能[7]。

一开始，人工智能能够比人类医生做出更精确的医学诊断，这一事实乍一看似乎令人惊讶，但人们早就认识到，在大多数情况下，统计推断都优于人类专家的临床判断。看到像沃森这样的人工智能是进行统计推断的理想方法，因此，使用计算机进行诊断可以拯救生命。

2.3 认知偏见：犯错是人的错

在统计推理中，人类专家比人工智能更缺乏能力的一个原因是上述(不幸的是，所有人都太过人性)倾向于高估自己的能力。这种倾向被称为过度自信倾向，只是众多有记载的认知偏差中的一种，它可能导致人类思维系统的错误。另一方面，人工智能可以避免认知偏差。原则上，对人工智能的预测越来越有信心，可能会导致对许多社会和政治挑战采取更合理、更有效的方法，前提是它们是安全的、根据可理解的标准进行的。这里的问题在于利用人工智能的优势，而不是在相应的系统中放弃人类的自主权。

3. 项目结论和展望

无论是现在还是过去，对新技术和基本有利技术的非理性恐惧都是普遍存在的。这种“技术恐惧症”也可能是沃森或无人驾驶汽车遭到质疑的原因之一。然而，对各种技术保持警惕并不总是不理性的。大多数技术都可以用来造福人类，但当他们落入坏人之手时，也可能是危险的，或者当人们对安全性和不可预见的副作用采取不够谨慎的时候。

这也适用于人工智能：无人驾驶汽车可以让我们的生活更轻松，拯救人类生命，但复杂的计算机算法也会让股市意外崩盘。尽管在不久的将来，特定领域的人工智能的风险似乎有限，但有长期的发展需要考虑；在不远的将来，人工智能在原则上可以构成生存威胁，类似于与生物技术相关的大流行风险[8]。

3.1 自动化和失业

鉴于最近在机器学习和机器人领域的成功，似乎只有时间问题，即使是需要高智商的复杂工作也能被机器全面接管。如果机器在许多领域比人类工人更快、更可靠、更便宜，这可能会导致劳动力市场被连根拔起，这是自工业革命以来从未出现过的。考恩、McAfee 和 Brynjolfsson 等经济学家认为，技术进步将进一步拉大收入差距，并可能导致大量人口收入下降和失业率上升。2013 年的一项分析得出结论，在 10 - 20 年内，美国有可能实现 47%的工作自动化[9]。要想实现自动化，最困难的工作是需要高水平的社会智力(如公关咨询)、创造力(如时装设计)和/或敏感和灵活的对象操作(如手术)。在这些领域，人工智能研究的现状仍远低于人类专家的水平。

李开复博士预测人工智能会带来大量失业。事实上，人工智能现在还远远不能取代大多数服务工作，因为机械手还远不如人手灵活。AlphaGo 可以击败世界上最好的围棋手，但是还是要靠一个工作人员的手把棋子放到棋盘上。据估计，机器人技术至少要 30 年才能取代人的清扫、空姐、保姆等工作。麦肯锡估计到 2050 年，现在一半的工作将被机器人取代。取代一半的工作，从经济上来讲，就是把生产率和每天的薪资提高一倍。这意味着在工作量不变的情况下，平均一个美国人每年收入大约十万美元。在这个收入水平，人们会减少工作吗？有

些人会每周工作 3~4 天，但是还会有很多人每周还是工作 5 天，然后把 10 万美元去消费更好的教育、医疗、旅游和其他高端服务。这就会产生巨大的服务业工作需求，完全可以吸收工作被人工智能取代的劳动力。

李开复博士的另一个预测是贫富差距会大幅上升，因为大公司将得益于人工智能的效率提升，从而赚取巨大的利润。问题是，历史上有很多革命性的技术也带来了巨大的效率提升和社会效益，但是发明或使用技术的公司并没有赚到巨大的利润。谁还记得哪个公司发明了蒸汽机、计算机或者机器人，并且赚了多少钱?一开始发明和使用新技术的公司能够赚不错的利润，但超额利润往往会很快消失在激烈的竞争中，最后消费者得到了低价和高质的产品服务，成为最大的得益者。的确，人工智能技术需要大量的数据，这有利于大公司。但是世界上有不少有大量数据的大公司，举例来说，在无人驾驶领域，Google, Tesla, Uber 还有几家中国公司都有足够的数据成为有力的竞争者。一般的规律是，只有要 3 个以上接近的竞争对手，就赚不到太多的利润，所以虽然中国和美国会成为人工智能的超级大国，但是其人工智能的公司不会有钱到可以影响国际关系。

因此，我认为与创新有关的工作在可预见的未来，将主要由人类完成。一方面，让电脑来做创新可能是太危险。另一方面，创新往往涉及审美和口味的判断。例如，如果任务是评估一段音乐、一部电影，或者一种新菜，人类可能永远比机器人更了解自己的需求。有人说，创新只需要少数天才，不需要大量人口。但这显然不符合历史趋势。人类在创新方面正在投入越来越多的资本和人力资源。这种趋势现在并没有放缓的迹象。将来会更多人具备能力和意愿来参与某种形式的创新活动，既包括高技能工作(例如人工智能编程)，也包括低技能工作(例如游戏测试和电影评论)。更多的人将具备这种能力，这部分归功于人工智能帮助他们提高分析能力。更多的人将具备这种意愿，因为参与创造是富有乐趣和满足感。即使是像电影评论这种轻松的工作，也是某种形式的创造性活动。

3.2 计算机自动化的优势与劣势

随着生产和服务行业的自动化，人们可能只会期望娱乐业继续存在；但在这里，我们也看到了巨大的变化。随着计算机图形技术的发展，新颖的娱乐技术，以及无数的智能手机应用都变得越来越便宜，视频游戏和互联网使用的吸引力正在上升。随着虚拟现实技术的日益普及，这些效应可能会被放大。随着这些变得越来越详细和现实，他们可能模糊了用户在现实和模拟之间的界限。

尽管娱乐产业提供人工智能的游戏化教学和学习材料,这也增加了风险，越来越多的年轻人会有困难完成他们的教育。

乌托邦和反乌托邦技术进步提高了社会生产力，反过来提高了平均生活水平。如果有更多的工作是通过机器进行的，这将为人类腾出时间来休闲和自我发展。然而，提高自动化程度的一个缺点可能是，生产力的提高伴随着社会不平等的加剧，因此，平均生活水平的上升与生活质量的上升并不一致。麻省理工学院(MIT)经济学教授埃里克•布林约尔松(Erik Brynjolfsson)等专家甚至担心，技术进步可能会使大多数人的生活变得更糟。

在竞争激烈的经济中，人工智能技术已经发展到许多工作都是由机器完成了，自动化人类工作的收入将会下降。如果没有监管，许多人的收入可能会低于生存水平。如果经济产出比有效再分配所需的工资增长更快，社会不平等可能会急剧上升。为了抵消这一发展，McAfee 和 Brynjolfsson 建议将某些工作限制在人类身上，应该得到补贴。

一些专家也对未来可能发生的情况发出警告，预测的变化更加剧烈。例如，经济学家罗宾•汉森(Robin Hanson)预计，在本世纪内，将有可能对人脑模拟，即所谓的“全脑仿真”(WBEs)：在虚拟现实中进行数字化运行，WBEs 可以是可再生的，而且可以（假设有足够的硬件可用）比生物大脑快很多倍，从而意味着劳动力效率的大幅提高。汉森预计，在这种情况下，WBEs 将会出现“人口爆炸”，因为他们可以被用作是高效的员工。汉森的推测是有争议的，不应该假设他们勾画出了最有可能的未来情景。目前在这一领域的研究，例如在 ETH Lausanne 的蓝

色大脑项目，仍然离第一个大脑的模拟还很远。然而，重要的是要记住硬件的发展与 WBEs 的可能性。如果汉生的设想成功，这将是非常合乎道德的。首先，许多人被复杂的模拟所取代；另一方面，有一个问题是，WBEs 是否会有非凡的意识和主观偏好。

4. 一般智力和超级智能

一般智力的衡量，是一个正常人在广泛的环境中达成目标的能力。如果他们的目标与我们的目标不一致，这种智能就会带来灾难性的风险。如果一个普通人的智力达到了超人的水平，它就变成了一个超级智能；也就是说，一种算法在任何方面都优于人类智力，包括科学创造力、“常识”和社会能力。请注意，这个定义留下了一个问题，即超级智能是否具有意识。

4.1 一般人工智能相对于人类的比较优势

人类是聪明的，他们是具有两条腿的“生物机器人”，拥有意识，并在数十亿年的进化过程中被开发出来。这些事实已经被证明，人工智能的创造可能并不那么困难，因为人工智能的研究可以更快、更有目标的进行，而不是进化(只有经过几代人的缓慢积累才会进步)。与人的生物大脑相比，计算机硬件有几个优点：基本的计算元素(现代微处理器)“火”的速度，比神经元快数百万倍；信号传输的速度了要快数百万倍；而且，一台计算机可以存储相当多的基本计算元素(一台超级计算机很容易占用整个工厂的地盘)。未来的数字智能将会比人类大脑更大的优势，与此相关的软件组件，例如，它很容易被修改和复制，这意味着任何时候都可以调用潜在的相关信息。在一些重要的领域，如能源效率，对纯粹物理损伤的恢复力，和优雅的退化，人工硬件仍然落后于人类的大脑。特别是，在信息处理水平上，热力学效率和复杂性降低之间仍然没有直接的联系[10]，但随着未来几十年计算机硬件的改进，这可能会改变。

鉴于这些比较优势和预测的硬件的快速改进，人类的智能很可能有一天会被机器所取代。重要的是要更精确地评估这种情况何时发生，以及这种情形的影响在哪里。

4.2 时间

人工智能领域的不同专家曾考虑过，第一台机器何时能达到人类智能水平的问题。根据一项引文索引，对 100 名最成功的人工智能专家进行了一项调查，结果显示，大多数人认为，人类水平的人工智能很可能在本世纪前半年内被开发出来。相信人类将在本世纪末创造出一个超级智能，只要技术进步没有大的挫折（由于全球灾难的结果），大多数专家也持同样的观点。但有一些专家相信，至少在 2040 年之前，将有人类智力水平的机。其他专家认为这一水平永远无法达到。即使一个人做出了一种保守的假设，即认为人类专家对他们的估计过于自信，在相关专家如此广泛的信心的情况下，将超级智能描述为仅仅是“科幻小说”仍然是不恰当的。

4.3 一般智力的目标

人工智能是否会有道德行为，也就是说，它是否有不与人类利益冲突的目标，是完全开放的。在原则上，人工智能可以遵循所有可能的目标。这将是一种错误的人格化，认为每一种超级智能都会对像人类的伦理问题产生兴趣。当我们建立人工智能时，我们也明确或含蓄地确立了它的目标。

这些主张有时会受到批评，理由是任何试图根据人类价值指导人工智能目标的尝试都将等同于“奴役”，因为我们的价值观将被强加于人工智能。然而，这种批评建立在一个误解上，因为人工智能在它被创造之前就已经存在了。创造智力的过程不可避免地决定了它的功能和目标。如果我们想要建立一个超级智能，那么我们，什么都没有，没有其他人，负责它的目标。此外，也不是说人工智能必须通过我们不可避免的目标经历任何形式的伤害。从道德上讲，被伤害的可能性需要意识，我们必须确保意识不是由超级智慧来实现的。父母不可避免

地会以一种非常相似的方式形成孩子“生物智能”的价值观和目标，但这显然并不意味着孩子们会以一种不道德的方式被“奴役”。恰恰相反：我们有最大的道德义务，将基本的道德价值观传授给我们的孩子。我们创造的人工智能也是如此。

计算机科学教授斯图尔特·罗素(Stuart Russell)警告说，道德目标的编程在技术层面上构成了一个巨大的挑战，无论是在技术层面上(在编程语言中，如何编写复杂的目标，以确保不会产生无法预见的后果)，还是在道德层面上(无论如何，目标是什么)。虽然超级智能的可能目标范围很广，但我们可以对他们所采取的行动做出一些可靠的声明。这些包括目标和自我保护，增加智力和资源积累。如果人工智能的目标被改变，这可能是消极的(甚至更多)，以实现其最初的目标，即毁灭人工智能本身。增加智能本质上是只是一个更广泛的环境中达到目标的能力,这开辟了一个所谓的智能爆炸的可能性,一个人工智能迅速经历了一个巨大的增长并自我完善。资源的积累和新技术的发现给人工智能带来了更多的动力，从而实现了更高的目标。如果新开发的超级智能的目标函数没有赋予有知觉的生命，那么它将会导致不计后果的死亡。

人们可能倾向于认为超级智能不会带来危险，因为它只是一台计算机，它可以完全拔掉插头。然而，从定义上来说，超级智能并不愚蠢；如果有任何可能被拔掉，超级智能就会像制造商希望的那样，开始表现自己，直到它发现如何将非自愿关闭的风险降到最低。超级情报也有可能绕过大银行和核武器武器库的安全系统，使用迄今为止未知的安全漏洞(所谓的“零日剥削”)，并以此来敲诈全球人口，迫使其合作。如前所述，在这种情况下，“回到最初的情况”是极不可能的。

4.4 什么是利害攸关的

在最好的情况下，超级智能可以解决无数的人类问题，帮助我们克服未来最伟大的科学、伦理、生态和经济挑战。然而，如果超智能的目标与人类或其他任何有知觉的生物的偏好不相容，那么它将成为一种前所未有的生存威胁，可能会比已知宇宙中任何先前的事件造成更多的支持。

4.5 合理的风险管理

在风险很高的决策情况下，以下原则是至关重要的:

(1) 只要有足够的机会赢/输，代价高昂的预防措施即使是低概率风险也值得付出代价。

(2) 当专家之间没有共识时，谨慎的谦虚是明智的。也就是说，你不应该对自己的观点的准确性有太多的信心。人工智能研究的风险是全球性的。如果人工智能研究人员在第一次尝试中未能将伦理目标转移到超级智能上，那么很可能就没有第二次机会了。估计人工智能研究的长期风险比气候变化的风险要大得多。然而，与气候变化相比，人工智能研究很少受到关注。在本文中，我们想要强调的是，将大量的资源投入人工智能安全研究中是更有价值的。如果这里讨论的场景有一个非无穷小的实际发生的机会，那么人工智能和与之相关的机会和风险应该是全球优先考虑的事情。良好的人工智能研究结果的概率可以通过一系列措施来最大化，包括以下几点：如果这里讨论的场景，也许很小，但超过无穷小的机会真的发生，那么人工智能和与它相关联的机遇和风险应该全球首位。通过以下措施，可以使人工智能研究取得良好成果的可能性最大化。

5. 人工意识

人类和许多非人类动物具有非凡的意识，具有主观性。他们有感官上的印象，一种(基本的或明显的)自我感觉，身体上受到伤害的痛苦经历，以及感受心理上的支持或快乐的能力。简而言之，他们是有知觉的生物。因此，在某种意义上，他们会受到伤害，这与他们自身的利益和观点息息相关。在人工智能的背景下，这引出了以下问题：机器的功能系统是否也可

能经历一个潜在的痛苦的“内在生命”？哲学家和认知科学家 Thomas Metzinger 提出了的四个问题，所有这些标准都适用于机器和动物:

(1) 工作意识。

(2) 一个非凡的 self-model。

(3) 项目在自我模型中，注册负值的能力(即，违反了主观偏好)。

(4) 透明度（即，感知感觉不可挽回地“真实”，从而迫使系统自我认同其意识自我模型的内容）。

前两个问题实际上必须要加以区别：首先，机器是否能够发展意识和对所有人的支持能力；其次，如果第一个问题的答案是肯定的，那么哪种类型的机器会有意识。还有，机器是否能够技术上发展意识，以及是否有能力为所有人提供支持;第二，如果第一个问题的答案是肯定的，那么哪种类型的机器会有意识。这两个问题正在被哲学家和人工智能专家研究。研究表明，第一个问题比第二个问题更容易回答。目前，专家们一致认为，机器可以在原则上有意识，而且至少在神经形态的计算机上是有可能的。这种计算机的硬件与生物大脑的功能相同。然而，识别哪种类型的机器(除了神经形态的计算机)可能有意识的问题要难得多。这个领域的科学共识不太明确。例如，有争议的是，纯粹的模拟(如蓝色大脑项目的模拟大脑)是否具有意识。虽然一些专家相信这是事实，但其他人不同意。

鉴于专家们的这种不确定性，采取一个谨慎的立场似乎是合理的：根据目前的知识，至少可以想象，许多足够复杂的计算机，包括非神经形态的计算机，可能是有知觉的。这些考虑具有深远的伦理影响。如果机器可以有意识，那么利用它们作为劳动力，并利用它们从事危险的工作，如拆除地雷或处理危险物品，将是不道德的。如果足够复杂的人工智能有一定的意识和主观偏好，那么人类和非人类动物使用的类似的伦理和法律安全措施将会得到满足。如果，比方说，蓝色大脑计划的虚拟大脑是为了获得意识，那么使用它(以及任何潜在的副本或“克隆”)来进行系统的研究，例如将其置于抑郁的环境中，将是非常有道德问题的。Metzinger 警告说，有意识的机器可能被滥用用于研究目的。此外，作为“二等公民”，他们可能缺乏法律权利，被当作可有可无的实验工具，所有这些都可能在机器的内在经验水平上产生负面影响。这一前景尤其令人担忧，因为可以想象，人工智能将会以如此巨大的数字出现，在最坏的情况下，可能会有天文数字的受害者，超过以往任何已知的灾难。

这些反乌托邦的场景指向了技术进步的一个重要暗示：即使我们只犯了“轻微的”伦理错误(例如错误地将某些计算机分类为无意识或道德上的不重要)，那么，由于历史上前所未有的技术力量，这可能会导致同样前所未有的灾难。只有承认可能机器意识的不确定性，我们才能开始在人工智能研究中采取适当的预防措施，从而避免上述任何可能的灾难。

6 结论

今天，我们正在见证新型人工智能技术的传播，其潜力惊人。目前无人驾驶汽车、沃森辅助医疗诊断和中国军用无人机的人工智能技术将在可预见的未来逐渐普及。至关重要的是，在这种情况发生之前，必须精心构建法律框架，以使这些技术的潜力能够安全地发展。人工智能技术领域取得的进展越快，对相关挑战的理性、有远见的方法就越迫切。因为政治和法律的进步往往滞后于技术的发展，因此，个人研究人员和直接参与任何进展的开发人员有一种特别大的责任。

然而，不幸的是，在没有风险分析的情况下，有强大的经济激励来推动新技术的发展。这些不利条件增加了我们逐渐失去控制人工智能技术及其使用的风险。在所有可能的层面上都应该避免这种情况，包括政治、研究本身，以及任何与这个问题相关的人。在最有利的轨道上指导人工智能发展，是一个基本前提。通过这种方式，它不仅能在少数专家中得到认可，而且能在广泛的公共话语中得到认可，成为我们这个时代的一个伟大挑战。

我们也不必担心现在的研发工程师会变穷或者没有工作。真正要担心的是程序员和数据

分析员的收入远高于清洁工的收入，正如李开复博士指出的那样。但是贫富差距加大不是一个新问题，解决办法还是要提高基本的社会保障和普及教育。人工智能技术带来的经济繁荣和教育智能化可以反而帮助政府解决贫富差距的问题。

总之，在不久的将来，会有更多的服务业工作机会出现，而创新相关的工作会不断增加。长远来看，创新不光是解决一个问题，更多的是探索未知的事物。人类对更多食物和住房的需求很容易饱和，然而人类总是有兴趣探索新的器具、新的故事、新的游戏，以及探索太空。如果人类不再有探索的欲望，那么人类文明将开始衰落。这是一个比“贫富差距”难得多的问题。

References

[1] Koomey, J. G., Berard, S., Sanchez, M., & Wong, H. (2011). Implications of Historical Trends in the Electrical Efficiency of Computing. IEEE Annals of the History of Computing, 33(3), 46–54.

[2] Brockman, J. (2015). What to Think About Machines That Think: Today’s Leading Thinkers on the Age of Machine Intelligence. Harper Perennial.

[3] Russell, S. (2015). Will They Make Us Better People? (http://edge.org/response-detail/26157)

[4] Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press. [5] BBC. (2015a). Stephen Hawking Warns Artificial Intelligence Could End Mankind. (http: / /www.bbc.com /news / technology-30290540)

[6] Harris, S. (2015). Can We Avoid a Digital Apocalypse? (https://edge.org/response-detail/26177)

[7] The Independent. (2014). Stephen Hawking: ‘Transcendence Looks at the Implications of Artificial Intelligence — But Are We Taking AI Seriously Enough?’ (http: / /www.independent. co. uk / news / science / stephen - hawking - transcendence-looks- at-the-implications- of- artificial-intelligence-- but- are-we- taking- ai- seriously-enough-9313474.html)

[8] The Guardian. (2014). Elon Musk Donates $10m to Keep Artificial Intelligence Good for Humanity. (http://www. theguardian.com/technology/2015/jan/16/elon-musk-donates-10m-to-artificial-intelligence-research)

[9]SBS. (2013). Artificial Irrelevance: The Robots Are Coming. (http://www.sbs.com.au/news/article/2012/07/18/ artificial-irrelevance-robots-are-coming)

[10]BBC. (2015b). Microsoft’s Bill Gates Insists AI Is a Threat. (http://www.bbc.com/news/31047780)