强健及有益人工智能研究重点

2015年1月11日

（翻译：王德宇）

执行摘要：在探索人工智能道路上取得的成功，将有可能为人类带来前所未有的益处。因此我们有必要研究如何将这种收益最大化，同时避免潜在缺陷的发生。本文举出了若干实例（绝对不要认为这些例子概括了全部可能性），来阐述这些具有价值的、为确保人工智能维持在强健可靠且有益而进行的研究。

1 人工智能现状

始有人工智能研究以来，人们探索了诸多问题及方法。但直到最近20年，人们的注意力才开始聚焦在搭建“智能体”（intelligent agent）的相关问题上。智能体是一种在特定环境中进行感知并产生行为的系统……

随着研究对象的能力范围从满足实验室研究环境到具备商业价值的技术，一种良性循环应运而生：即便是小小的一点性能改进，也值得大笔经费的投入，用以推动相关研究的开展……

人工智能研究所取得的进展，让我们开始不单单关注其能力，还需探索如何将其社会效益最大化。这样的考量，促成了诸如AAAI 2008-09人工智能未来发展领导人级会议，以及其他针对人工智能的影响所发起的社会项目……本文的研究必定是跨学科的，因其涉及了社会与人工智能两方面，包括经济、法律、哲学，以及计算机安全性、各个时期、各类分支的人工智能技术。本文重点探讨如何使人工智能更具社会良性效益，更具保证正面影响的可靠性，即我们的人工智能系统需要按照我们的意图运转。

2 短期研究重点

2.1 优化人工智能的经济影响

人工智能在工业产业中的成功应用，包括制造、信息化服务等，已经证明其对经济越来越大的冲击，尽管人们很难对这种冲击的准确性质进行定义，也难以界定人工智能与其他信息化技术带来的不同影响。许多经济学家与计算科学学家共同意识到，在人工智能领域，还有很多值得研究的内容，包括如何将经济效益最大化、同时降低不利影响，如更多的不公、更高失业率等[50，12，25，26，71，52，48]。这些考量促进了一系列相关研究等开展，包括经济学、心理学等。下面列出一些例子，对上述内容进行不全面的解释。

1. 人才市场预测

我们将如何预期，各类工种在何时、以何顺序被自动化设备取代？这一自动化过程，将如何影响技能稍欠工人、创意人士、及其他各类信息工作者的薪资待遇？有观点认为人工智能或将极大程度地增加人类的总体财富。然而，自动化程度的增加，可能会使收入的分布更趋指数化。

2. 对市场其他方面的冲击

大部分经济部门，包括金融、保险、精算，以及很多消费市场，都可能遭受人工智能技术运用于学习、建模、预测行为所造成的冲击……

3. 处理不利影响的对策

哪些政策有助于自动化越来越高的社会变得更加繁荣？举例来说，Brynjolfsson和McAfee[12]探索了一系列激励劳动力密集产业、促进运用人工智能所产生的财富来支撑失业人口的政策。另外，我们还需关注诸如教育培训改革、学徒计划、劳力需求体系项目、最低工资法规及税收结构的调整、社会安全网等干预政策有何优缺点。历史上有很多无需为温饱而工作的社会群体的例子，包括古代的贵族、当今卡塔尔等国家特定阶层的公民等。怎样的社会结构及其他因素可以决定这些阶层人民是否富裕？失业并非意味着休假，同时失业与幸福感丧失、自我怀疑、孤独感等有着很深的联系[34，19]；理解怎样的政策及社会常态可以打破这些关联，将大大有助于改善生活质量。关于基本收入标准等议题的经验研究及理论研究，将理清我们所面临的不同选择。

4. 经济指标

人均GDP等经济指标，有可能无法准确反映深度人工智能化及自动化对经济体所带来益处与损害。这使得这些指标不利于政策制定者来参考[50]。对于能够更好反映问题对经济指标对研究，将有利于决策的制定。

2.2 法律及伦理研究

高度智能和自动化系统的开发，会引发有关法律和伦理的重大议题。这些议题的最终结果，将影响人工智能技术的生产者和消费者。这些议题涉及法律、职业道德、哲学论题等方面，同时需要来自计算科学、法律、政策制定、伦理学等方面等专家共同探讨。

1. 针对自动驾驶车辆的权责划分与法律规范

……

2. 机器伦理

……

3. 全自动武器

……

4. 隐私问题

……

5. 职业道德

……

从政策角度来说，人工智能（同其他任何强大等新技术一样），在带来新的机遇的同时，也会引发诸多弊端。合理的应对政策，能够确保人类在享有新技术带来的益处的同时，将风险降至最低。关于政策的问题包括：

1. 哪些领域的政策值得研究？

2. 评估政策的标准有哪些？可能的标准包括合规的可验证性、执法可行性、降低风险的能力、避免开发不利技术的能力、适用性、随情况转变的可适应性等。

2.3 针对强健人工智能的计算科学研究

随着自动化系统在社会中越来越普遍的应用，它们能够稳健地按照我们地意愿执行工作，变得愈发重要。自动驾驶汽车、自动贸易系统、自动武器等项目的开发，激发了人们针对高可靠性系统开发的兴趣。这样的系统可以确保具有极高的稳健性。Weld和Etzioni曾声称，除非我们找到了可信任的方法来确保自动系统的安全，社会是不会接受这些自主运转体的[88]。人工智能系统未按照意愿执行功能的不同失效方式，对应了有关稳健性研究的不同领域：

1. 正确性验证：如何证明一个系统满足特定的性能需求。（我是否正确地搭建了该系统？）

2. 效度：如何确保一个符合规格的系统，不会执行错误的行为，引发不当后果。（我是否搭建了正确的、正当的系统？）

3. 安全性：如何避免未授权个体或组织，进行不当操纵。

4. 控制：在人工智能系统启动，开始运转后，如何有效进行必要的人工控制。（好吧，如果我搭建的系统出错了，我可否修正它？）

2.3.1 正确性验证

……

2.3.2 效度

……

2.3.3 安全性

……

2.3.4 控制

……

3 长期研究重点

一些人工智能研究者经常讨论的长期目标是，为系统开发如人类般从经验中进行广泛学习的能力，同时使系统可以超越人类完成大多数认知任务的能力，因而对社会产生巨大影响。在可预见对未来，如果这样的努力哪怕是有一线希望得以实现，那么除了前文中提到的内容之外，还需要更多的相关研究，确保人工智能具有稳健而有利地为人类服务地能力。

不同研究者对于我们能够成功研发稳健有利系统的机率的估计，具有很大差异。但基于以往的各种推论预测，鲜有研究者能够非常自信地说这个希望是渺茫的。极少数例子表明，否定的论断并不会阻止技术的发展。Ernest Rutherford，这位可以被称作当时最伟大的核物理学家，在1933年曾号称核能的利用完全是空想（作者注：原话为“原子分裂所释放的能量非常微弱，想要指望通过这种原子转变来提供能量的人完全是在空想。”[58]），皇家天文学家Richard Woolley也曾在1956年将星际旅行描述为“完全是胡扯”[62]。退一步说，只要人工智能技术得以发展的机率大于可忽略的程度，我们就有理由为其投入足够多的资源。好比我们会愿意为房屋保险而进行充分的投资，即便发生火灾的概率极其微小。

3.1 正确性验证

……

3.2 效度

……

3.3 安全性

……

3.4 控制

……

4 结论

总的来说，在探索人工智能道路上取得的成功，将有可能为人类带来前所未有的益处。因此我们有必要研究如何将这种收益最大化，同时避免潜在缺陷的发生。本文举出了若干实例（绝对不要认为这些例子概括了全部可能性），来阐述这些具有价值的、为确保人工智能保持稳健可靠且有利人类权益地情况下运转而进行的研究。

参考文献（见原文）