文章编号：1672-8785(2020)07-0025-05

人工智能可控性探究

张明柱　薛沛祥　于　淼　陈庆磊

（中电科仪器仪表有限公司，山东 青岛　266000）

摘　要：由当前人工智能的某些不可控案例出发，联想到人工智能的脆弱性与不确定性，试探性地提出了提高人工智能可控性的方法。主要分为五个部分：第一部分首先简述人工智能的提出与发展；第二部分介绍聊天机器人以及其他人工智能产品的具体不可控案例；第三部分主要探讨人工智能技术脆弱性和不确定性的来源；第四部分提出提高人工智能可控性的几点建议，包括增强人工智能的场景感知能力、在系统架构中部署可控节点以及通过可控节点健全监管评估体系；第五部分描述人工智能技术在光电等领域的展望。

关键词：人工智能；脆弱性；不确定性；可控性

中图分类号：TH7　　**文献标志码：**A　　**DOI：**10.3969/j.issn.1672-8785.2020.07.005

Research on Controllability of Artificial Intelligence

ZHANG Ming-zhu，XUE Pei-xiang，YU Miao，CHEN Qing-lei

(*China Electronics Technology Instruments Co.，Ltd.，Qingdao 266000，China*)

Abstract: Starting from some uncontrollable cases of current artificial intelligence (AI), the fragility and uncertainty of artificial intelligence are associated, and the methods of improving the controllability of artificial intelligence are tentatively proposed. The article is mainly divided into five parts. The first part briefly describes the proposal and development of AI. The second part introduces the specific uncontrollable cases of chat robots and other AI products. The third part mainly discusses the sources of AI technology fragility and uncertainty. The fourth part puts forward some suggestions for improving the controllability of AI, including enhancing the scenario awareness ability of AI, deploying controllable nodes in the system architecture and improving the supervision and evaluation system through the controllable nodes. The fifth part describes the prospects of AI technology in the field of optoelectronics.

Key words: artificial intelligence; fragility; uncertainty; controllability

0 引言

人工智能(Artificial Intelligence，AI)技术正在全球范围内兴起，影响面广，颠覆性强，其在商用领域的应用极大地增强了公共服务与城市管理的水平，而且在军事领域的应用也提高了装备的智能化水平。例如，美国某驱逐舰的舰船作战系统可以在所有舰内人员尚未反应的情况下自动识别来袭导弹并对其进行拦截。

收稿日期：2020-06-03

作者简介：张明柱(1987-)，男，山东青岛人，工程师，硕士，主要研究方向为微波教学与智能测试。

E-mail：zmzjob@foxmail.com

正是由于 AI 技术的蓬勃发展，其脆弱性与不确定性亦会在未来展现，比如聊天机器人在评论中发布激进言论、在埃航失事事件中存在人机操作冲突等。这不禁引起我们的联想与思考，即随着 AI 发展，机器与算法必然会涉及到人类人身和经济财产的安全。因此人们需要研究如何提高 AI 技术的可控性，以避免因错误判断而造成巨大的负面效益。本文就此展开详细分析，并探究提高 AI 可控性的方法，最后对 AI 在红外成像和智能测试等领域的应用作出展望。

1 AI 的提出与发展

自 1956 年被提出以来，AI 技术经历了各界学者的不断探讨与发展。传统的主流研究方法有三大类：符号主义、联结主义与行为主义[1]。其中，符号主义认为智能是对符号的计算和推理过程；联结主义认为人的智能由人脑的生理结构与工作模式决定，在用计算机实现智能时要着重模拟人脑结构；行为主义则阐述智能取决于对外界环境的感知与行为。

经过六十多年的发展，AI 技术已成为了能引领未来的战略性技术。各大领域不断地将 AI 技术引入自身体系，以提高智能性与便捷性。例如，文献[2]中的避雷器检测引入了红外图像处理方法，并利用 AI 相关的机器视觉和数据分类算法实现了故障避雷器的自动识别；文献[3]则提出了一种基于 UNet 深度学习模型的遥感信息提取算法，提高了遥感图像地物信息自动提取的精确性。但是 AI 发展的脆弱性与不确定性同样也带来了新的挑战。如何引领 AI 安全、可控地发展必然会成为一个迫在眉睫的问题。

2 AI 的可控性问题

2.1 聊天机器人的不当言论

“微软小冰”是微软(亚洲)互联网工程院基于云计算、情感计算框架等多种综合技术的跨平台 AI 产品，是当前娱乐型聊天机器人的经典代表之一。除此之外，亦存在过如 Tay、BaBy Q 等众多聊天机器人。当前聊天机器人可以利用深度学习和数据挖掘等技术，通过在线学习实现对当前新闻热点的跟踪和评论。在未来的规划中，机器人可以控制多种设备并可实现上百种场景操作以及个性化的私人定制服务。然而随着当前聊天机器人的实际应用，可控性问题愈发突出，特别是互联网中数据的多样化与某些网友有意或无意的误导，意外频频发生。如何避免聊天机器人发表过激言论，其深层问题体现在如何提高 AI 的可控性上。

2.2 埃航空难的人机操作冲突

埃塞俄比亚当地时间 2019 年 3 月 10 日，埃塞俄比亚航空公司一架全新的波音 737-MAX8 飞机坠毁。飞机上来自 35 个国家的 157 名人员全部遇难。根据事后找到的驾驶舱录音记录与飞行数据，飞机坠毁前驾驶舱出现了人机操作冲突。机动特性增强系统(Maneuvering Characteristics Augmentation System，MCAS)持续命令客机机头下坠，飞行员则试图拉高飞机，然而 MCAS 的高权限最终导致了飞机坠毁失事惨剧的发生。

未来 AI 必然会在人类的各种社会活动中得到广泛应用，比如汽车自动驾驶、基于机器视觉的智能检测等。但伴随生产力提升而来的脆弱性和不确定性也应当引起我们的反思。

3 AI 的脆弱性与不确定性

2018 年 6 月，美国安全中心发表了《人工智能：决策者需知》报告。该报告指出，随着 AI 应用的扩张，人类将处于前所未有的全球革命之中，AI 的某些弱点同样会对未来的经济、军事、国家安全等领域产生重要影响，其中 AI 的脆弱性与不确定性是两个重要的方面[4-5]。

3.1 AI 的脆弱性

当前 AI 的应用领域主要包括数据分类、异常探测、预测推理和任务优化等四个方面[6]。其中，数据分类和异常探测都是基于大数据利用新工具和新理论进行类型判别的；预测推理则基于数据提取特征变化趋势，在状态

空间内进行建模分析；任务优化利用算法与数据对任务进行重新整合，并在各层节点间进行评估搜索[7-9]。由此可知，AI 与数据息息相关。当前的 AI 在一个封闭环境中时可能表现良好，但当面对一个开放环境时，AI 却异常脆弱。一旦无法归类或判别新出现的数据，AI 可能就立刻无法思考。

3.2　AI 的不确定性

确定性是对真实世界的近似刻画，不确定性则辨证反映了真实世界。AI 技术中的不确定性主要存在于以下四个步骤中：信息获取的不确定性、认知的不确定性、形成知识的不确定性以及决策结果的不确定性[10-11]。

AI 在获取信息时，常会受到各种客观因素或偶然因素的干扰，导致收集到的数据不完备甚至相互矛盾。例如，聊天机器人可以通过与网友互动进行学习，而某些网友的态度或者言论却比较极端。AI 在认知过程中需要对知识或数据进行建模或学习，而且该过程中同样存在不确定性。比如，现在常用的神经网络在建模时隐含层的层数选择会对建模产生巨大的影响，进而形成知识的不确定性。当前数据建模完成后，会得出相关的理论与结果。由于系统本身的稳定程度不同、数据选取的特征不同，有些结果并不被用户关心，例如动态贝叶斯网络的拓扑结构在推理学习时就有较大的不确定性。AI 做出决策时，其不确定性依旧存在。当前的 AI 并不能充分掌控问题和环境。不同的智能算法根据不同的出发点可能做出迥异的决策，例如，博弈论中的纳什均衡，如果设计不当，有些决策甚至会与用户的初衷相背离。

4　AI 的可控性

如图 1(a)所示，常见的 AI 技术没有为使用者提供足够的可控机制。本文借鉴统一建模语言(Unified Modeling Language，UML)中的场景活动图，从系统的顶层架构入手，建议设计反馈子系统，同时预留人工可控节点。整体如图 1(b)所示，提高 AI 的可控性，应对可能存在的安全风险。

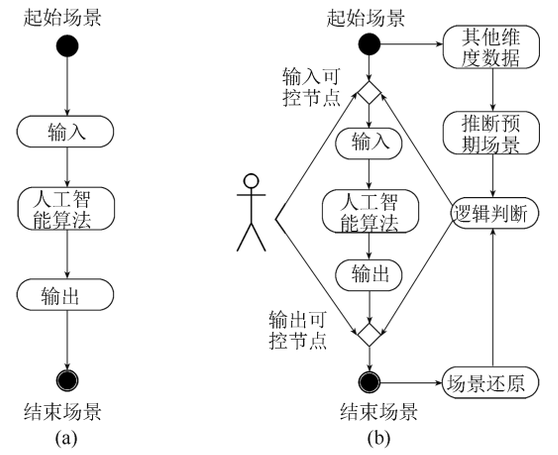

图 1　提高 AI 可控性的场景示例图

4.1　增强 AI 的场景感知能力

增强 AI 体系中的场景感知能力，提高数据驱动的可靠度[11]。在埃航事故发生后，波音公司对 MCAS 进行了升级。他们将原先读取单一迎角传感器数据改为读取双迎角传感器数据，以避免某单一迎角传感器出错导致 MCAS 做出错误决策。前文所述的埃航事故正是由于传感器出现失误导致飞机飞行异常。该项升级引出了一个问题：当一个 AI 系统决策出现失误时，仅靠增加传感器数目和提高系统输入数据的准确性，能否避免 AI 决策再次出错？结合前文内容可知，提高可控性并非如此简单。与此同时，还需提高算法的复杂度，降低 AI 决策的权限和权重等。这些做法都应用了图 1(a)所示的 AI 决策流程——AI 系统从输入端口获得起始场景的数据，经算法计算之后，将决策输出。考虑到一个整机系统的复杂性，如果每次应用 AI 时都需要额外冗余的传感器、计算空间和决策权衡，那么整个系统一定会非常复杂且造价高昂。

本文认为一个高可控性的 AI 系统应充分利用数据融合技术来实现场景感知和语言感知等多维度感知方法。如图 1(b)右侧所示，AI 利用直接输入的起始场景信息作出决策，然后需进行场景还原；与此同时，需要采集起始场景的某一直接关联信息(如起始场景的原因、结果等纵向信息，或者人为操作预期等横向信

息），并推断预期场景；再将还原后的场景信息与预期场景信息进行交叉对比，并在逻辑判断后对算法进行改进或者自我控制开闭，形成反馈闭环。需注意，该系统的其它维度信息可作为其他系统的直接输入，实现数据（传感网）的复用，做到同一起始场景下数据源的分离，从而降低系统的脆弱性。实际场景与预期场景的差异也有助于提高 AI 性能和增强系统的鲁棒性。最终实现从数据到决策、从决策到预期的 AI 认知计算模型。

4.2 在系统架构中部署可控节点

利用前沿理论研究成果提高 AI 的自适应能力，同时部署人工可控节点。虽然目前高级机器学习理论已有部分成果，但仍需考虑在利用当前理论实现信息获取、认知、推理和决策的过程中充满不确定性[12]。因此软件及物理架构应同时考虑自适应与可控性，关键环节中应增加控制节点，以防止错误决策的发生。由于不确定性存在于 AI 的整个决策过程中，将可控节点置于整个 AI 系统中并不具备可操作性。本文建议至少在 AI 系统与外部的交界处设置可控节点，并将其作为输入，为人工控制预留出接口。如图 1(b)左侧所示，其中设置了输入可控节点与输出可控节点。输入可控节点决定了 AI 当前的信息输入状态，在输入信息缺失时可有效进行人工干预；而输出可控节点则决定了 AI 的决策输出状态，有效避免了当系统出现错误决策时人工无法修正的状态。当然，在预期场景与实际场景不一致时系统应能自我关闭。可控节点从人机协作的角度表明了 AI 技术是为人类服务的。

4.3 通过可控节点健全监管评估体系

可通过可控节点部署状况建立 AI 安全监管与评估体系。AI 技术是一种用于扩展人类能力的技术，它不仅提高了人类的体力上限，更提高了人力的脑力上限。AI 的正确决策固然使人心安，但是作出错误判断必然会产生极大的负面效益。737-MAX8 客机坠毁前的“人机搏斗”正是机器判断与人类判断相左时的情景。

AI 自提出后就被应用在人类生活的各个领域，特别在数据推理方面表现惊人。2016 年 3 月，谷歌公司的 AIAlphaGo 机器人在围棋领域击败了世界顶级棋手李世石，引发了极大关注。我国大批学者对 AI 理论及前景进行了深入探讨[13]。但同时也不能忽略伴随 AI 而来的问题。如何正确使用 AI，AI 又会对个人和集体安全产生何种影响？因此有必要建立完善透明的 AI 监管体系，并构建 AI 安全监测预警机制[14-15]。由于可控节点可以有效管控 AI 系统中的数据流动，可通过评估 AI 系统的可控节点数目与部署位置的方法，开发系统性的测试方式，落实防控手段，从而确保 AI 在安全可控的范围内发展[16]。

5 展望

AI 技术是一种具有巨大社会效益与经济效益的革命性通用技术，其数据分类、异常探测、预测推理和任务优化等应用均可与光电、射频等领域交叉结合，产生众多的新产品和新技术，如文献[17]所述的红外目标建模方法以及文献[18]所述的水彩笔油墨红外光谱模式识别。由此可见，AI 的发展可以有效提高生产力的上限。需注意的是，在充分利用 AI 技术的同时，只有对 AI 的脆弱性加以规避和对不确定性加以限制，才能避免背离初衷，从而更好地服务社会。

参考文献

[1] 宝达理. 人工智能引发的问题研究 [D]. 北京：北京交通大学，2018.

[2] 卢彬，朱海峰，谷振富，等. 基于红外图像的避雷器故障检测方法 [J]. **红外**，2018，**39**(1)：19-23.

[3] 陈睿敏，孙胜利. 基于深度学习的红外遥感信息自动提取[J]. **红外**，2017，**38**(8)：37-43.

[4] Cai S S，Xue X D，Wu L W. Artificial Intelligence and Human Intelligence－On Human-Computer Competition from the Five-Level Theory of Cognitive Science[J]. *Contemporary Social Sciences*，

2017，**46**(4)：140-155.

[5] 闫志明，唐夏夏，秦旋，等. 教育人工智能(EAI)的内涵、关键技术与应用趋势——美国《为人工智能的未来做好准备》和《国家人工智能研发战略规划》报告解析 [J]. **远程教育杂志**，2017，**35**(1)：26-35.

[6] 杨晓庆. 计算机多功能智能化的可控性研究 [J]. **信息与电脑(理论版)**，2014，**6**(8)：56.

[7] 孙志军，薛磊，许阳明，等. 深度学习研究综述 [J]. **计算机应用研究**，2012，**29**(8)：2806-2810.

[8] 杨旭. 计算机网络信息安全技术研究[D]. 南京：南京理工大学，2008.

[9] 王光宏，蒋平. 数据挖掘综述 [J]. **同济大学学报(自然科学版)**，2004，**32**(2)：246-252.

[10] 潘道华. 基于计算机模拟的不确定性推理研究 [D]. 哈尔滨：黑龙江科技学院，2009.

[11] 张昕. 人工智能中的不确定性问题研究[D]. 长沙：国防科学技术大学，2012.

[12] 马修由·谢勒著. 曹建峰，李金磊译. 监管人工智能系统：风险、挑战、能力和策略 [J]. **信息安全与通信保密**，2017，**39**(3)：45-71.

[13] Qiu J. Research and Development of Artificial Intelligence in China [J]. *National Science Review*，2016，**3**(4)：538-541.

[14] 孙柏林. 美国新的人工智能报告及其对我们的启示[J]. **自动化技术与应用**，2017，**36**(10)：1-7.

[15] 陈伟光. 关于人工智能治理问题的若干思考 [J]. **人民论坛·学术前沿**，2017，**6**(20)：48-55.

[16] 何哲. 人工智能时代的政府适应与转型 [J]. **领导决策信息**，2017，**24**(5)：15.

[17] 苗壮，张湧，李伟华. 基于双重对抗自编码的红外目标建模方法 [J/OL]. **光学学报**，http://kns.cnki.net/kcms/detail/31.1252.o4.20200602.0854.012.html，2020.

[18] 王晓宾，马枭，王新承. 基于人工神经网络的水彩笔油墨红外光谱模式识别 [J/OL]. **激光与光电子学进展**，http://kns.cnki.net/kcms/detail/31.1690.TN.20200601.1754.167.html，2020.