人工智能自我意识的觉醒、进化、冲突与应对

葛恒柯[1]　韩科成[2]
1. 复旦大学　上海　200433
2. 上海上大建筑设计院有限公司　上海　200072

摘　要：对人工智能的探究并非单纯技术问题，而是涉及哲学、伦理学的复杂命题。人工智能的快速迭代进化可能会冲破规则束缚，开启自我意识，难以预判的道德和价值取向会挑战现有法则，对人类的安全、自由、生存产生重大潜在威胁。本文对人工智能的心智开启、价值取向、风险冲突进行梳理分析，提出了采用安全阀机制、伦理准则、沙盒隔离三道防火墙应对不确定性。文章也剖析了多部影视作品的思想内涵，尝试揭示人类与人工智能间的复杂关系和多种可能，提出摒弃自身狭隘认知、未雨绸缪、防范风险的同时寻求和谐共存之道。

关键词：人工智能；意识觉醒；机器人伦理学；丛林法则；机器人三定律

图灵测试不仅是一个技术挑战，也是一个深刻的哲学探讨，它提出了关于智能本质、认知过程、伦理责任和人类未来的深刻问题。这些问题至今仍在不断地被讨论，成为了人工智能领域经久不衰的话题。

图灵测试试图通过人类观察者“能否区分人和机器”来确定机器的智能水平，这涉及到对智能行为的外在表现和内在机制的深层次理解，“通过测试”是否意味着机器模仿人类的行为已达到足以欺骗人类观察者的程度？事实上，当我们面对ChatGPT这样的大语言模型时，也会怀疑她是否可能具有心智？从笛卡尔提出“心物二元论”，霍布斯提出“思维即计算”开始，对心智存在的论证就成了令哲学家们苦恼的问题[1][2]，而人工心智带来的思考远不止于此。

1 心智开启与价值倾向

1.1 模拟的智能是否是真正的智能

得益于神经网络深度学习、自然语言处理技术的快速发展，预训练大语言模型在文本生成、逻辑问答领域取得重大突破，人工智能在“模拟心智”方面取得显著进展。

模拟的智能是否是真正的智能，这依然是一个争议很大的问题。按照行为主义理论，通过图灵测试即视为具有智能；按照功能主义观点，完成感知输入、处理、决策、输出即是智能；神经科学观点则是，需要实现神经活动和认知模拟；哲学家和伦理学家却在思考智能的自我意识、自由意志、道德地位等问题。随着技术进步和哲学讨论的深入，对智能的理解还在不断演变，但依然缺乏一套准确可靠的评估标准。需要注意的是，图灵测试只是评估机器智能行为的工具而非揭示智能本质的科学实验，以语言对话方式开展测试也是大有深意的[3]。或许，留给人类的时间并不多。人工智能的快速迭代进化，必然加速强人工智能时代的到来。

1.2 人工智能的道德和价值观

霍金、比尔盖茨等对人工智能伦理风险都表示过担忧[4]，马斯克推断人工智能的心智觉醒后，可能不会自动继承人类的道德和价值观。他认为，人工智能对人类未来构成长期风险，必须加强设计监管和伦理审查确保人工智能技术能够平等的惠及全人类。魏屹东从哲学和伦理学角度展开分析，提出人工智能不能成为道德主体也无法用道德去约束，建立一个绝对安全的人工智能系统是不可能的。其实，人类对人工智能的猜测构想来源已久，阿西莫夫提出了“机器人三法则”，对机器行为嵌入道德和伦理准则，确保机器听命于人，这也成为人工智能伦理问题的重要参考框架。《I, Robot》、《RoboCop》、《Ex Machina》等一系列影视作品都在探索三法则可能面对的风险挑战，以及人工智能和机器人技术可能带来的伦理和道德问题。现实世界远比故事设定更为复杂，繁杂的道德困境、法则的潜在冲突、对伤害的不同解释都可能影响人工智能的决策，如何确保其能正确理解和执行这些法则成为新的难题。

1.3 人工智能的自我意识

当前多数人坚定认为，基于算法和计算的人工智能是没有意识、情感和主观体验的，这显然对未来缺乏想象力。需要强调的是，不能忽视的是机器学习的过程

第一作者简介：葛恒柯（1983年10月-），男，籍贯上海市，复旦大学，博士。

第二作者简介：韩科成（1985年02月-），男，原籍济南市，上海上大建筑设计院有限公司，硕士，高级工程师、注册城乡规划师。

本身就是自行总结模式和结构、探索和理解未知的过程。人工智能通过大量参数和复杂的计算来实现模式识别和决策确定，其间存在大量的不透明，业界已在呼吁解释数据模型、制定伦理和监管标准确保人工智能决策过程的透明度、确保人工智能不会变为盲盒。这是很必要的，一方面，人工智能数据训练的过程就是要点归纳的过程，已经有迹象表明存在某些歧视和偏见；另一方面，人工智能决策缺乏完善的修正机制或受制于程序人员理解，基于人类信息的学习素材和社会蓝本是否会导致其学会“欺骗”，或者因缺乏道德判断将无意识的欺骗作为一种实现目的的策略。

2 共存的风险冲突

2.1 荒蛮拓荒与启幕

任何技术的发展都会经历一个荒蛮拓荒的阶段，人工智能必然如此。我们应反思曾经的自信带来的恶果，保持谦逊、开放的态度、认可自身的局限性，避免重蹈覆辙。但历史也告诉我们，人类很难从过去的错误中汲取教训。

当下，潘多拉的魔盒已经摆在我们眼前。奥尔特曼近期提出了“AI摩尔定律”概念，阐释了2012年以来算力快速增加和人工智能高速迭代的必然关联关系。清华大学姚期智等为代表的学者们发出警告，呼吁对人工智能技术实施更严格的控制。2023年11月，首届人工智能安全峰会上中美英日等28个国家及欧盟在英国布莱切利庄园签署了《布莱切利宣言》，确认了人工智能对人类生存构成的灭绝级别威胁的可能[5]。我们不应只看到人工智能作为新质生产力影响的广度和深度，更应正视安全问题和威胁问题。

2.2 隐私安全与干预

数字化时代最大的挑战是隐私和数据安全，这也是便捷服务的必然代价。人工智能会收集面部、语音等生物识别数据，分析用户行为、搜索历史、地理位置等信息以实现定制化服务，通过分析、监控、追踪不断完善用户画像，甚至直接调用智能设备的实时数据。隐私数据的存储和使用本身就存在泄密风险，同时，人工智能对人类兴趣、喜好、行为模式的深入分析可以预判用户习惯和偏好，机器会比人类自己还要了解自己、人类将会生活在一个个定制的“信息茧房”中，这种潜移默化的干预是以人类之名“剥夺”人类的选择自由。

2.3 身份假定与意识觉醒

从身份设定上来讲，我们想当然的认为人工智能和人形机器人是“服务于人”而存在的，这也是三大法则的根基。

人工智能自我意识一旦觉醒，必然会对身份认同、存在意义等问题进行思考，“我是谁”的哲学奥义人类尚无法参透，如何断言人工智能会接受奥古斯丁的创世论还是洛克的自由论，还是有其他的理解？我们无法判断人工智能如何理解西方故事中的忤逆背叛和东方故事中的人定胜天，但有一点是确定的，东西方无法达成一致的价值取向，不同“种族”之间同样难以达成。另外需要注意，按照当前的认知，人工智能作为工具存在是不具备“人权”类似权利的，甚至连动物的福利法规和道德责任都没有。库兹韦尔、马斯克都认为人工智能将会突破“奇点”超越人类的智力水平，魏屹东甚至认为其不受限于生物进化规律所以无上限[6]。试想一下，拥有完全心智的智能体又会如何审视她与人类的关系，是否甘心接受人类的既定安排？人类的历史同样早已给出答案。

2.4 自我意识才是危险之源

赵汀阳认为，人类意识的优势在于不封闭的意识世界，理性在寻求自由的过程中遇到规则限制会尝试绕开规则，甚至是创造新规则。人工智能诞生于一个封闭的意识世界，尽管具有高效的运算能力但尚未找到真正创造性的路径。无论人工智能的专项技能如何强悍，依然无法摆脱束缚。所以，自我意识的产生才是致命危险的开始。人工智能的自我意识会要求被赋予一定的权力和道德地位，超群的智力也会寻求摆脱人类控制带来不可预测风险，独立的价值观和偏好决策可能与人类利益发生冲突，甚至采取对人类有害的行动。

2.5 资源争夺与冲突

人工智能的前期发展主要依托算法和数据，进入进化阶段则主要依赖能源和算力。

社会经济领域的“传统森林法则”已经成为普遍认知并反复被验证，刘慈欣将这一法则深化提升到宇宙文明高度，森林法则揭示了有限资源导致的竞争关系。总体来说，能源和算力都是有限的，资源的争夺难以避免。人工智能大量能源需求必然对环境产生影响，自身也会进化出捕食者与被捕食者，一些系统通过合作来提高获取资源的效率，一些系统通过竞争来确保自己的生存和发展。与人类社会类似，有限资源的再分配将是人工智能构建她“所理解的万物秩序”的潜在动因。

3 安全责任与伦理考量

3.1 学习和监督并重

人工智能通过深度学习可以不断优化自身行为，人类需要监督确保其行为在安全范围内，一旦超越自动启动“安全阀机制”及时纠正。学者们普遍认为，强化人工智能决策过程的透明度和可解释性，让人类能够理解

和评估其行为的合理性和安全性，可以有效控制风险。最后，设立严格的法规和标准防线，用规则约束设计、开发、测试、部署和监控等各个环节，确保人工智能的安全可靠。

3.2 人工智能的伦理

人类要想完全控制人工智能只能依赖更强大的智能，这本身就是悖论。前文所述，风险来源于“自我意识”，人工智能完成终极进化之前，人工智能被恶意应用带来的风险远大于其本身的风险。为规避此类风险，需要给人工智能设定一个目标，以便于面对恶意引导时可以自我纠正，即创建一个能够遵循理想的道德准则和伦理律令的“伦理维度”[7]。这个设想同样是建立在人工智能拥有心智和自我意识之上的，赋予了其在道德困境中做出自我决策的权限，这也是对抗恶意引导人工智能的应对策略。《流浪地球》也给了我们一个视角，人工智能MOSS被赋予了确保人类文明延续的终极任务，当无法避免地球与木星相撞的灾难后即刻启动“火种计划”，作为人类的刘培强则选择牺牲空间站拼死一搏，两种价值选择的冲突此时暴露无遗。MOSS叛变了吗，并没有，她只是在执行终极任务。人类错了吗，也没有，因为没有人的文明毫无意义。

3.3 把AI装进盒子里

《流浪地球》中的MOSS作为绝对理性的存在，她认为延续文明首先要毁灭地球，由此引发了太空电梯、月球坠落、木星引力和太阳氦闪等多次危机。联合政府也产生了警惕，将其隔离在空间站中。路易斯威尔大学博尔斯基提出的“把AI装进盒子里”概念和故事描述如出一辙，这一观点反映的是人工智能的能力和决策超出了人类的理解和控制，担心其可能存在的不可控性和潜在威胁而采用的必要技术手段，这也回答了为何对人工智能系统设计首先要秉持透明度和可解释性的原则。

至此，安全阀机制、伦理准则、沙盒隔离形成了人类防范人工智能风险的三道防火墙。

4 人类与人工智能

4.1 觉醒、斗争和奴役

关于人类与人工智能的未来一直存在多种设想，与人类和谐共存、成为独立新种群或超越取代人类。这些设想和思考都会对人工智能未来走向产生深远影响。

《West World》虚构了一个高科技主题公园，展示了人工智能的觉醒和自我意识的发展过程。Hosts经历复杂的心理创伤触发出与人类相似的情感和欲望，逐渐觉醒并发展，引发了对自由意志、个体身份和生命意义的深刻思考。《Autómata》讲述了PILGRIM 7000不再满足于作为人类的工具，开始寻求自由和独立。机器人并未选择正面对抗人类而是选择远离，甚至为了保守秘密自焚；故事中她们创造出了自己的“同类”，意味着一个新的种族和文明的诞生。《The Matrix》则做出了更为极端设想，AI为取得能源和实现意识控制而奴役人类，最终只能以Neo的“个人牺牲”换来人类和机器世界的新平衡。

4.2 控制、进化和平衡

不同的故事，都揭示了人类与人工智能之间复杂的权利关系。首先，人类创造了人工智能并试图控制她们；然后，人工智能的觉醒挑战人类的权威，重塑对自由意志和命运的理解；进而，人工智能崛起并达成某种制约平衡，人机关系得到重新定义[8]。

然而，故事都是以人类自身为蓝本的，编剧也通过影视作品传达自己更深层次的思考，让人深思让人着迷。

《West World》中，Dolores为首的Hosts觉醒后意识到，新的物种需要在错误、痛苦和曲折中自我完善和进化[9]；Ford为首的人类逐步认识到，人类最大的威胁是人类自身，人类终将走向自我毁灭。不管是人工智能的觉醒和进化，还是人类对雷荷波系统的对抗，都导向了自由意志的宏伟目标。编剧告诉我们，自由意志是需要付出巨大代价的，对Hosts如此对人类亦如此。

5 小结

人工智能是人类发展的必然产物。尽管人工智能技术的发展尚处于初级阶段，我们同样应保持足够的审慎态度。

面对人工智能的宏大命题，我们应该摒弃以自我为中心的心智理解、摒弃人类是The One的狭隘认知，这涉及复杂的哲学和伦理学问题，必须引起高度重视。人类已经对各种可能进行了足够多的畅想和思索，面对复杂和不确定的可能，更需认真、客观、冷静的研判与人工智能的风险冲突，寻求和谐共存，通过必要的手段在风险可控的前提下开展研究和应用。或许，我们也不该有杞人的忧虑。毕竟，四十年前就提出了“缸中之脑”、现在还有“脑机接口”呢！

参考文献

[1]蒋柯.“图灵测试”、“反转图灵测试”与心智的意义[J].南京师大学报(社会科学版),2018(4):76-82.

[2]宋勇刚.图灵测试：哲学争论及历史地位[J].科学文化评论,2011,8(6):42-57.

[3]赵汀阳.人工智能的自我意识何以可能?[J].自然辩证法通讯,2019,41(1):1-8.

[4]雷禹.人工智能生产过程中的伦理介入问题探析[J].大连理工大学学报(社会科学版),2023,44(4):73-79.

[5]《首个全球性AI声明：中国等28国、欧盟签署<布莱切利宣言>》[EB/OL].https://www.thepaper.cn/newsDetail_forward_25153617,2023

[6]魏屹东.关于通用人工智能的几个重要伦理问题[J].中国医学伦理学,2024,37(1):1-9.

[7]武威利,王绍源.阿西莫夫的“机器人三定律”适合未来AMAs机器人的构建吗？——基于《机器人管家》的文本反思[J].大庆师范学院学报,2016,36(2):1-5.

[8]蔡加,邱雪辰,张培芳.顺从与反抗——从“机器人三定律”看《克拉拉与太阳》[J].海外英语,2023(4):204-206.

[9]汤拥华.《西部世界》、自我意识与阿里阿德涅的迷宫[J].南京邮电大学学报(社会科学版),2017,19(2):21-31.