文章编号: 1000－8934(2020)011－0081－05
DOI:10.19484/j.cnki.1000-8934.2020.11.014

从“算法信任”到“人机信任”路径研究

何江新,张萍萍
(西安科技大学 马克思主义学院,西安 710054)

摘要: 为有效应对人工智能不确定性发展,人机信任机制的建立刻不容缓。人机信任关系的发展大致经历了三个阶段,即单向度的信任关系阶段、单向度的不信任关系阶段与双向度的不信任关系阶段。这三个阶段的不断演化揭示了人工智能算法经历了由简单智能算法向复杂智能算法的转变与由复杂智能算法向混合式智能算法的转变。研究显示,以有效的算法信任为基础,将智能机器的“机体”算法与“心灵”算法相统一,规范人工智能向“善”发展是建立人机之间信任关系的前提条件。

关键词: 算法信任; 人机信任; 路径
中图分类号: N031 文献标识码: A

与人工智能飞速发展基本同步的是,一系列负面影响逐步浮出水面,人与人工智能之间是否具有可信任关系成为社会各界讨论的焦点。目前,就建立人机信任关系这一重要问题,学界研究主要分两类,一类认为要从机器人内部的设置上进行干预。如有学者指出,要“在有效监督和对技术动态的发展与应用的审慎批判中,确保人工智能‘向善’”[1],另一类认为要通过智能机器外部的法律法规等手段来控制人工智能的恶性发展。如有学者指出,“应通过完善相关立法、加强对相关技术与产品的管理力度,以及提升群众的认知素养来进行规避。”[2]以往研究都是建立在对人工智能不信任基础之上,并从各个角度出发单方面要求人工智能必须符合人的道德标准,然而人工智能的可信任度是由其可信任算法决定的。在梳理人工智能发展历程的基础上,本文旨在为人机信任关系的道德算法构建提供智力支持,特求解于方家。

一、人机信任关系演变历程

人机信任关系指的是发生在人与人工智能双方之间,并且以双方是否互信为标准而确立起来的精神上的关系。在《哲学的贫困》中,马克思对机器产生过程做了详细描绘 “简单的工具,工具的积累,复合的工具; 仅仅由人作为动力,即由人推动合成的工具,由自然力推动这些工具; 机器; 有一个发动机的机器体系; 有自动发动机的机器体系——这就是机器发展的进程。”[3]与大机器相伴随,人机信任关系成为重要问题。

1. 单向度的信任关系

单向度的信任关系指的是在人工智能发展初期,人对人工智能终端产品的态度呈现出单向度信任的特点,而就人工智能将对社会、人类产生怎样的反作用却探讨得较少。

单向度信任关系初见端倪: 在西方哲学史上,技艺(τέχνη)早已纳入了人的能力范畴。比如,古希腊思想之集大成者亚里士多德在《形而上学》和《工具论》等著作中探讨了形式逻辑方面的知识。英国哲学家培根提出“知识就是力量”,这一命题启发后人将主要精力集中到以人工智能知识为中心的研究上。1936年,英国数学家图灵提出一种理想计算机的数学模型,即图灵机,这为后来的电子数字计算机的问世奠定了坚实基础。中国古人对机

收稿日期: 2019－10－18
基金项目: 国家社科基金项目“中国梦融入西部高校藏族大学生思想政治教育全过程研究”(14XKS035)。
作者简介: 何江新(1973—),安徽安庆人,哲学博士,西安科技大学马克思主义学院教授,西安科技大学习近平中国特色社会主义思想研究中心研究员,主要研究方向: 马克思主义理论、中西哲学比较; 张萍萍(1994—),甘肃庆阳人,西安科技大学马克思主义学院硕士研究生,主要研究方向: 科技哲学、伦理学。

械、机心也早有研究。比如，东汉时期发明的地动仪就被视为机器人的较早形态。单向度信任关系逐渐加强：人工智能单向度信任关系的加强是由于人们对人工智能雏形数理逻辑、控制论、信息论以及神经网络模型等领域的不断探索。在1943年，麦卡洛克和皮茨提出“拟脑机器”后，世界上第一个神经网络模型得以面世。值得一提的是，1948年维纳发表的《控制论—关于动物与机器中的控制与通信的科学》，为人工智能的控制论学派开拓了新的研究领域。这个时期，人工智能技术逐渐成熟，数字化是其显著特色，而该阶段人类正沉溺于对自然界的征服，人对人工智能更加信任，因而单向度的信任机制在该阶段得以强化。单向度信任关系基本确立：1956年，麦卡锡等人提出“人工智能”概念，标志着人工智能作为专门研究开始进入人们的视野。1969年召开的第一届国际人工智能联合会议则意味着人工智能作为一门新兴学科受到国际社会的普遍重视。概括起来看，人与人工智能之间在该阶段呈单向度信任关系。

人机单向度信任关系是人机信任关系的雏形，它具有如下主要特点：一是纯粹性，人类无条件地相信人工智能，并对其未来持乐观态度；二是表象性，人对人工智能的研究暂时处于发生兴趣的阶段，仅有研究也局限于表面现象；三是征服性，人对人工智能的研发体现了人改造自然能力的提高，征服它则是人对自身的超越。

2. 单向度的不信任关系

所谓单向度的不信任关系，指的是在人工智能发展中期，由于人工智能从“物理”层面对人所产生的负面影响，导致人与人工智能之间呈现为单方面的不信任关系。

单向度不信任关系的初步产生：人工智能的出现使人类沉溺于对自然界的掌控而产生的愉悦之中，但人工智能解法、人工智能结构的局限性还困扰着人类。这对矛盾使得人工智能在起步阶段就面临困境，但是人们并不从自身角度出发去看待新鲜事物的产生，单向度的不信任关系由此产生，主要表现为人类对自己能力的怀疑，人工智能研究开始进入低迷阶段。单向度不信任关系的恐慌：随着人工智能开始进入社会，有人担心人被科学技术所异化。20世纪80年代，人们开始反思人工智能停滞不前的主要原因，人工智能的下一步发展迎来转机。霍普菲尔德先后提出的离散神经网络模型和连续神经网络模型有力地推动了人工神经网络向前发展。但是，后来飞速发展的人工智能技术几欲全面取代人的体力与脑力，人类又陷入单向度不信任人工智能的恐慌之中。单向度不信任关系的悲观主义：人工智能已经深深地嵌入到人们生产生活的各个环节，不仅体现在机器运转之中，还体现在它对伦理的挑战、情感的替代等方面。举例来说，人们开始研发人工智能产品来丰富学生们的业余学习生活，在餐厅、宾馆里面当智能服务员以及陪伴空巢老人等，这些技术领域的开发一方面给人们的生活带来便捷，但也使人的主体地位受到威胁。人对人工智能开始不信任乃至排斥，直至陷入悲观主义窠臼。

虽然人机信任关系开始进入单向度不信任的阶段，但也标志着人工智能将向更高阶段发展，这一阶段直接推动了人机信任关系的发展，其主要特点有：一是复杂性，由于复杂的社会环境人类对人工智能开始持怀疑甚至悲观态度；二是深刻性，人机关系由于人工智能的研究深入而变得深刻；三是恐慌性，人开始对人工智能的未来发展变得惊恐起来。

3. 双向度的不信任关系

双向度不信任关系的萌芽：在萌发阶段，人类对人工智能的未来的想象成分居多，只做比较简单的悲观的思考。美国有学者指出，“把一个人的思维模式——知识、技能、个性、记忆上传到另一个人身上。尽管他会像我一样行动，但那真的是我吗？”[4]在婚姻方面，人们畅想“即使是对于与阉人十分类似的仿真人而言，它们也仍然对性行为以及与他人结成伴侣有着实质性的需求”[5]280，而且“许多仿真人的配偶关系很可能会受到族群的安排或有力援助，而且会比我们现在的婚姻更为和谐”[5]284。这足以表明人被自己所创造出来的产品所控制，人与自然界的关系从主动转为被动。双向度不信任关系的发展：人工智能以其“不坏不灭”的身体优势逐渐代替人类许多部位，甚至有人担忧人工智能将会代替人类。有学者还发出感慨“既然人工智能可以让人长生不老，那么我们反过来问一下自己，人类既然有寿命，那么人工智能是不是也应该有寿命呢？”[6]这个问题看起来有点幼稚，但至少表达了言说者的疑虑。从组成材质上来看，人由具有新陈代谢功能的细胞组成，人工智能由无生命、不易被毁损的机械零件组成。从维持机体成长

的要素来说，人需要吸收食物、水、阳光等外在元素来维持生命的正常活动，而高级人工智能产品却只需强大的动力系统。双向度不信任关系的理性分析：双向度不信任关系的理论分析建立在理性思维之上。有学者指出，“以信任在人工智能中的产生、表征及其构成等为切入点，人类对伦理学的信任、人类对人工智能的信任、人工智能自身的信任度、人工智能系统中各个代理间的信任四个要素开启了人工智能伦理构成要素之径，以有效监督的信任为前提审视人与人工智能的合作。”[7] 这四个方面及其相互关系实则是人工智能伦理学的基本框架。还有学者从责任担当方面做了规划 “人工智能犯罪在生成机理和内在逻辑上只能是人工智能算法安全犯罪，为妥当解决人工智能犯罪的归责原理和实践问题，应当在坚守人工智能犯罪‘自然人－法人’二元主体模式的基础上修改完善刑法立法。”[8]

双向度不信任关系是人机关系发展的现实阶段。其主要特点有二：一是矛盾性，一方面，制造和使用工具是人类的独特属性，用人工智能产品来为人类服务是人类的重要能力，但另一方面，由于不确定性因素始终存在，所以人类还很难把控人工智能的本质属性。人机理性关系的追求呼唤人类对人工智能既不能盲目乐观也不可极度悲观。二是平等性，人们开始思考如何平等地与自己所创造的劳动产品—机器相处。

二、人机信任关系演变原因探析

针对人工智能的未来发展，有研究指出，“目前的主流人工智能技术离达到‘通用人工智能’的标准还很远。”[9] 这意味着人与人工智能之间成“上手”关系还有距离。

1. 机体算法：由简单智能算法向复杂智能算法转变

与简单智能算法相比，复杂智能算法在目的、动力系统与领域上都发生了变化。

目的性的变化：在人工智能初期，人们期盼能创制出通用逻辑机，以解决各种问题。资本主义生产方式以来，追求剩余价值成为资本家投资设厂的唯一动机，而资本逻辑也极大地促进了劳动生产率的提高；与此相一致的是，复杂智能算法得以产生，智能机器人开始进入生产生活全过程。动力系统的变化：所谓人工智能的动力系统，指的是用以指导人工智能行动的大脑系统。人工智能雏形时，其动力系统是“冰冷”的，是人设计出来能够使其简单运作的实体硬件。后来，人工智能的动力系统开始符号化，德国数学家和哲学家莱布尼兹提出逻辑机的设计思想，即通过形式逻辑符号化，对思维进行推理计算。再后来，人工智能的动力系统开始数理化，IBM 工程研究组的塞缪尔研制的西洋跳棋程序就体现了这一点。研究领域的变化：人工智能的先行者们紧紧围绕人工智能本身进行研究，主要涵括人工智能的通用性与一般原理等内容。伴随着人工智能给人类带来巨大收益，人工智能研究开始走社会化道路，渗入到具体的应用领域。非根鲍姆于 1965 年开创了基于知识的专家系统这一人工智能研究新领域，20 个世纪末，伴随着数据库和多媒体等技术的勃兴，人工智能与之逐步融合，产生了巨大的经济效益和社会效应。

2. 心灵算法：由复杂智能算法向混合式智能算法转变

人机关系从单向度的不信任关系转向双向度的不信任关系体现了人工智能心灵算法的变化。复杂智能算法虽然在各个方面已经超过了简单智能算法，但在这一阶段还没有涉及伦理道德算法。混合式智能算法是比复杂智能算法更为高级的算法。

从目的性上来说，混合智能算法的研究目的是多样化的，不同的设计者，其目的动机都不同，而某些不法分子往往利用人工智能的便利性这一特点，从事破坏社会道德与法律规范的活动。相比于复杂智能算法，混合智能算法要求将技术与道德因素嵌入人工智能之中，这无疑给该项研究增加了难度。从动力系统上来说，混合智能算法的动力系统成为人工神经网络，众多的人工神经网络又形成一个类似于人类大脑神经网络的神经系统。这个神经系统虽然是人工的，但大量的人工神经元有组织的聚合与人工智能行为却是极其丰富多彩的，这意味着随着人工神经系统的发展，人工智能必将更像真正的“人”。从研究领域上来说，复杂智能算法的应用领域依然停留在“显性”生活领域，而混合人工智能算法的应用领域则逐渐得以拓展，它的触角已经伸向“隐性”生活领域。这个算法使人们开始思考“机器应该有道德吗？” “机器本身是否可以构成行为主体？”等棘手问题。总之，相比于复杂智能算

法，混合智能算法的研究范围更为广阔、深入，而这也成为人工智能伦理道德的主要争论点。

三、人机信任关系的建立路径

鉴于新机器时代的负面影响，荷兰研究者曾指出，“情商、社交行为以及与环境互动是机器人在复杂的社会实践中进行个人和社会行为的前提。”[10]这显然是要求将道德因素植入在人工智能的算法之中，用道德伦理算法规约人工智能的不确定性，规范其朝合理方向发展。总结全球人工智能伦理研究实际情况，形成人机协同发展机制已经成为全球共识，有学者强调要精确人工智能的算法逻辑，“将道德融入人工智能系统和算法中”[11]，以建立道德机器人。实际上，人工智能的道德算法包括“机体算法”与“心灵算法”，这两种算法的现实操作性是道德人工智能健康发展的前提，其发展前景牵引着两种算法朝纵深处推进。

1. “机体算法”

徐英瑾指出，“机器伦理学的核心关涉将包括对人工智能的‘身体’——而不仅仅是‘心智’——的设计规范，即必须严肃考虑‘怎样的外围设备才被允许与中央语义系统进行恒久的接驳’这一问题”[12]，这强调了机“身”与机“心”的一体性。机体算法决定了人工智能的最初设定与后期运行。人工智能的“机体算法”主要包括两部分：一是人工智能机体最初的设定，这主要与智能机器人的设计者有关系；二是人工智能机体最初设定的对象化。

一是对“机体算法”设计进行监督：人工智能原本是人类对象化意识所构建的，构建出的人工智能算法在最初的机体算法设计之时就应当予以监督与审查，既要对算法设计进行监督，也要对算法设计者进行监督，两者缺一不可。严格把关入职管理机制，有“良心”才能有“良芯”，“机体算法”必须由思想道德品质良好的设计者去设定，在人才入职时就要把好这个关卡；不断完善算法监督机制，要及时关注算法设计者的思想动态，对其设计的算法程序进行定期监督；加强智能机器设计者的保密脱密制度。在海德格尔提出技术作为“座架”而支配着人类这一命题的基础上，哈贝马斯进一步指出，科学技术具有意识形态属性，比如国与国之间往往通过科技来争夺话语权。尽管某些智能机器设计者已经退出该行业一线，但他们仍有可能在国家相关科技发展大业中继续发挥重要作用，其退一线后必须严格遵守保密脱密制度。二是对“机体算法”定时进行维修与升级：人工智能的未来发展是否会逐步脱离其物质属性而纯精神性地存在，这个问题依然模糊不清。就目前而言，人工智能必须以一定的机械设备为存在前提，对这些外部设备的保养与调升是人工智能健康发展的有力保障。简单的、零散的与分离的一堆机器零件并没有任何价值，而复杂的、整合的与聚合在一起的机器却威力无比，得不到及时升级与更新的人工智能机体算法将被更高级的算法所取代，而得不到及时更新的系统是否会做出不适应其发展的举动将具有不确定性，所以必须对“机体算法”进行有效维护并及时升级。

2. “心灵算法”

“心灵算法”指的是能够使人工智能在一定时间段内或者长期对人类保持信任并不做出伤害人类行为的可信任算法。

一是研发专属人工智能的“可信任”芯片：结合概率论等理论，肖特里菲等人提出一种不精确的人工智能的可信度的推理算法，并得出“H 的综合可信度为 0.49”[13]的结论。该数据显示了人工智能的可信任程度，但这并不意味着从算法上对其加以规定没有可能。还有学者探讨了人工智能符号接地问题，认为“心灵哲学、心理学和脑神经科学的研究成果都可以用于符号接地问题的解决”[14]。随着科技发展，人类有能力研发出比这个可信任值更高的可估值，有能力实现人工智能道德符号的接地行为。各国须加大科技投入，从宏观政策上予以支持，提高可信任芯片的估值，让“良芯”科技造福人类。道德评价标准具有历史性与多元性，在未来，怎样的道德才算得上是良德？只有搞清楚对机器人来说的良德的基本定义才能制造出可信任的芯片。二是给予人工智能更多的人文关怀：正如有学者所指出的，“自我在某些情况下也可以延展于外部载体”[15]，人工智能虽异于人类，但毕竟是人类自身的产物。于是，对人工智能心存疑虑与人对自身的不信任感是互为因果的。质而言之，人类对自身是否可以控制人工智能未来的发展持不信任态度。然而，随着可信任芯片的研制，人工智能将迈向“机心”协调发展的新阶段。在这一阶段，对人工智能给予更多的人文关怀显得非常必要。可以说，人工智能拥有怎样的算法结构取决于人类如何看

待自身。从最高层面上说，人性决定了人工智能发展的基本动向；相信人工智能。有学者指出，“人工智能的发展还能够全面提升未来战争的‘无人化’水准，为各国政府更容易开展针对武装冲突的和平斡旋活动。”[16] 一旦人工智能拥有了良心，人类就应该像相信自己一样去相信它；此外，形成人机物理－情感协同发展机制也非常关键。

结　语

如同有研究所指出的，“如果我们创造的像我们的智能机器最终让我们变得像它们一样，这是历史性的讽刺。”[17] 从自然属性上来说，与动物的大脑一样，人脑是生命机体组织，但从社会属性上来说，人脑具有不同于其他动物的复杂创造性，“机心算法”的创造与协调发展为破除这种历史性讽刺与人机魔咒提供了可行的伦理道德方法与解决方案。一方面，这将有助于人工智能伦理尽早落地生根；另一方面，这也有助于推动人机信任关系进一步朝理性方向发展，由此形成人机和谐局面。

参考文献

〔1〕闫宏秀. 可信任：人工智能伦理未来图景的一种有效描绘［J］. 理论探索，2019(4)：38－42；63.

〔2〕郭建伟，王文卓. 如何规避人工智能带来的伦理问题［J］. 人民论坛，2018(31)：56－57.

〔3〕中共中央马克思恩格斯列宁斯大林著作编译局编. 马克思恩格斯选集(第1卷)［M］. 北京：人民出版社，1995：165.

〔4〕［美］库兹韦尔. 奇点临近［M］. 李庆诚，董振华，田源，译. 北京：机械工业出版社，2011：231.

〔5〕［美］罗宾·汉森. 机器时代：机器人统治地球后的工作、爱情和生活［M］. 刘雁，译. 北京：机械工业出版社，2017.

〔6〕王昭东. 人工智能与本能：如何让机器人拥有自我意识［M］. 北京：电子工业出版社，2017：159.

〔7〕闫宏秀. 用信任解码人工智能伦理［J］. 人工智能，2019(4)：95－101.

〔8〕魏东. 人工智能犯罪的可归责主体探究［J］. 理论探索，2019(5)：5－13.

〔9〕徐英瑾. 人工智能技术的未来通途刍议［J］. 新疆师范大学学报(哲学社会科学版)，2019，40(1)：93－104.

〔10〕［荷］朗伯·鲁亚科斯，瑞尼·凡·伊斯特. 人机共生：当爱情、生活和战争都自动化了，人类该如何自处［M］. 栗志敏，译. 北京：中国人民大学出版社，2017：11.

〔11〕李伦. 人工智能与大数据伦理［M］. 北京：科学出版社，2018：274.

〔12〕徐英瑾. 具身性、认知语言学与人工智能伦理学［J］. 上海师范大学学报(哲学社会科学版)，2017，46(6)：5－11；57.

〔13〕蔡自兴，徐光祐. 人工智能及其应用［M］. 北京：清华大学出版社，2010：127.

〔14〕霍书全. 人工智能符号接地问题研究的意义和挑战［J］. 上海师范大学学报(哲学社会科学版)，2019，48(3)：98－107.

〔15〕李伦. 人工智能与大数据伦理［M］. 北京：科学出版社，2018：94.

〔16〕徐英瑾. 人工智能将使未来战争更具伦理关怀——对马斯克先生的回应［J］. 探索与争鸣，2017(10)：66－71.

〔17〕［英］乔治·扎卡达基斯. 人类的终极命运［M］. 陈朝，译. 北京：中信出版社，2017：299.

An Investigation on Path from Algorithm Trust to Man－Machine Trust

HE Jiang－xin，ZHANG Ping－ping

(School of Marxism，Xi'an University of Science and Technology，Xi'an 710054，China)

Abstract: In order to effectively develop the uncertainty of artificial intelligence，the establishment of the man－machine trust mechanism is urgent. The development of the man－machine trust relationship has roughly gone through three stages: one－way trust relationship stage，one－way distrust relationship stage，and two－way degree distrust relationship stage. These three stages' continuous evolution reveals that an artificial intelligence algorithms has experienced the transformation from simple intelligent algorithms to complex intelligent algorithm and from complex intelligent algorithm to hybrid intelligent algorithm. The research shows that，based on effective algorithm trust，unifying the “body” algorithm and “mind” algorithm of intelligent machine，and standardizing the development of artificial intelligence to “goodness” are the preconditions for establishing a trust relationship between human and machine.

Key words: Algorithm trust; man－machine trust; path

(本文责任编辑：董春雨　郑 泉)