

从系统功利主义看人工智能的价值原则冲突与应对

薄 闻

广西民族大学科技史与科技文化研究院 530006

摘要：近年来，人工智能的伦理反思主要围绕“人工智能责任主体”、“人工智能社会伦理风险”等方面展开。国内外学者对人工智能的态度有着二元对立的特点，一方面期待人工智能给人类社会更多的社会福祉，另一方面也对人工智能所带来的伦理风险有所警惕。无论是技术伦理设计、算法数据风险、还是社会伦理道德主体，其根源都在于道德哲学上功利主义原则于道义原则的冲突。融合功利主义原则于道义原则的系统功利主义在应对人工智能伦理原则冲突上具有较强的解释力，是应对人工智能伦理原则冲突的有效依据。

关键词：系统功利主义；人工智能伦理；原则冲突

1 引言

人工智能作为计算机科学的一个分支，该领域的研究包括智能机器人、语言识别、图像识别、面部识别和大数据处理等，人工智能的主要目标就是让机器替代人类智能从而完成一些复杂的工作^[1]。诞生于 20 世纪 50 年代的人工智能技术，随着社会和科技的不断发展，已经成为了本世纪最热门、高端的技术之一。目前的人工智能技术在艺术、医疗、教育、金融、军事等多个领域都有广泛的涉及，也衍生出了新的名词，比如元宇宙、万物互联等。人工智能作为一种可以改变人类生活方式的技术，必将带领人类走向科技发展的新时代。

人工智能在各个领域掀起波浪的同时，人们必须更加理智地思考人工智能带来的问题。想让人工智能更好地促进人类社会的物质文明和精神文明，除需了解人工智能的发展历史、应用领域外，还须从科技伦理学的角度去探讨人工智能的本质，并设计人工智能伦理冲突的解决方案。

系统功利主义是融合功利主义与道义论的理论体系。现有对人工智能的伦理反思与讨论本质上是对人工智能伦理原则冲突与应对的讨论。因此，如何解决人工智能伦理冲突是现阶段不可回避的问题。

2 人工智能伦理反思的发展历程

从人工智能技术出现的那一天开始，关于人工智能伦理问题的思考就开始了。在过去的 50 年里，人工智能技术进展比较顺利，相应的人工智能伦理反思也随着技术的发展而展开，人工智能伦理反思大概有三个阶段：萌芽阶段（2005 年前）、兴起阶段（2005-2018 年）、发展及繁荣阶段（2018-至今）。^[2]

国外对人工智能伦理的研究起步较早，“机器人三定律”是由美国作家艾萨克·阿西莫夫于 1942 年在其著作《I, Robot》中提出的，以保证机器人的发展不会危害到人类生存发展，此后“机器人三定律”成为了学术界公认的研发原则。1950 年，在图灵测试进行之后，人工智能之父艾伦·图灵大胆预言了未来将会出现有真正智能的机器人。1956 年，美国达特茅斯学院举办了历史上第一次人工智能研讨会，被认为是人工智能产生的标志。BonnefonJF 等（2016）对自动驾驶汽车的社会困境进行了讨论，指出目前的科学技术无法设计道德价值与个人利益平衡的算法，对于一些人工智能伦理问题，需要在人工智能产品量产前做出合理的决策。^[3] O'neilC（2016）认为算法不是中立的，人类的意志会通过设计的算法体现出来，并利用大数据进行杀熟等手段，这对于底层人民是不公平，作者认为应该对人工智能进行监管，并出台相关法律。^[4] AllenC（2005）并指出科技伦理的一个主要目标是要像道德代理人那样，设计出能够使其成为道德代理人的人工代理人。人工道德的中间目标是建立人工智能系统，对某种活动的价值、道德规范和正当性做出相应的反应。^[5] 国外学者对 AI 的研究时间长，研究程度深，涉及面广。有参考价值，可供我国人工智能伦理研究参考。

我国人工智能伦理学发展较晚，虽然现阶段研究结果比国

外少，但是也形成了一定的研究氛围，尤其是在哲学领域。^[6] 王东浩（2013）认为，机器人的人格化特征是随着自主趋势的发展而逐渐扩大的，与之相对应的，应该负起一定的责任。^[7] 段伟文（2017）指出：人工智能体的拟主体性赋予人工智能以拟伦理的独特作用。基于对合成智能与人工劳动者的价值审度、对智能的合理性及其与人的存在价值的冲突等价值诉求的反思，我们可以提出寻求算法决策与算法权力的公正性，呼唤更透明、更可理解、更可追责的智能系统。^[8] 周玄、赵建超（2022）为人工智能目前体现在四个方面：人的边界模糊，人的存在异化，人的自由异化，人的种族终结。^[9]

在人工智能技术发展的历史可以看出，科技的进步与社会伦理的进步应该相一致，社会伦理应该引导科学技术向正确的方向发展，这样才能促使科技更好地为人类服务。

3 人工智能伦理的理论选择

AI 伦理学的一个主要目的是增加人类社会对于 AI 的伦理学维度。人工智能伦理的研究目前的重点是找到确保人工智能符合人类社会行为规范的伦理理论，并使这种道德约束在人工智能之间相互传递，这是人工智能领域适合的伦理理论。解决 AI 道德的种种问题。

3.1 人工智能伦理中的道义原则

道义论起源于希腊语 Deon，意为“应该的”、“适当的”，是一种以人的行为动机为标准，不以结果来判断人的行为对错，以人的行为出发点是否为善、是否合乎道德准则为中心，要求人们必须按照某种道德原则展开行动的道德学说。^[10] 道义论的代表人物康德认为当人对道义原则有所感知时，人身上就会有一种义务感出现，然后人就会根据这种义务感去展开活动，也就是说人的行动是受道义原则所制约的，所展开的各种活动也是受到道义原则命令的。

在对康德道义论后续的研究中，有一种较有说服力的理论表现为“牺牲者中心”。^[11] 这种理论认为作为系统中的每个人，都有可能被系统中的其他人所伤害，都可能成为潜在的受害者，与此同时 AI 的发展，也许会让人类成为潜在的受害者。^[12] 对人工智能伦理问题来说，人工智能与人类共同构成了一个系统，人工智能发展过程中只要人类的每个个体本身得到了保障，那就说明人工智能无害于人，人工智能发展是合乎道德标准、符合道义原则的。反之，如果人工智能的发展可能使人类成为潜在牺牲者，那它就违反了道义原则，不具备道义合法性。从道义论的道义原则来看，如果人工智能具备自我意识且对人类存在潜在危害时，那么人工智能的出发点就是不正当的，道义原则就会阻止这种不正当的发生。虽然康德道义论在理论上是可行的，但是却忽略了实践过程中可能存在的各种问题。首先，康德强调，行止的起点要向善，即道义是纯粹性的，是唯心主义的，把体验和理性的内在联系割裂开来；其次，强调行为动机的善是片面的，忽略了行为客观结果；最后，缺乏对道义论的现实思考，使道德实践成为了一种形式主义。^[13] 所以，对于人工智能的伦理思考不能只是以“善”为出发，还需考虑结

果,我们必须引入功利论。

3.2 人工智能伦理中的功利原则

对功利原则的论述最早由边沁提出,功利原则又称“最大幸福原则”,本质上是追求大多数人的幸福。与道义原则不同的是,功利原则是一种结果论,即评价一个行为善恶与否,是根据该行为的结果是否符合绝大多数人的利益为标准。这种原则需要计算不同情况下的不同结果,也符合人工智能伦理所追求的道德计算。^[14]功利论与道义论是相反的理论,道义论强调的是行为出发点善恶与否,功利论强调的是行为导致的结果是否符合大多数人的利益。功利论因为过分强调结果,所以也是片面的理论。首先,对于功德量的多少没有固定的评价体系,比如一个富翁捐了100万元和一个乞丐捐了100元,不能说富翁比乞丐更高尚,因为爱心不能用财富来衡量;其次,功利主义只是单向寻求利益,但个人利益与集体利益之间的关系十分复杂,功利主义没有考虑行为主体的情感;最后,功利主义往往是以牺牲个人利益换取集体利益,但我们不能以损害他人正当利益换取利益最大化。

对于社交软件的普及,有些人认为社交软件可以让人们不再支付高昂的花费就可以交流,更可以与千里之外的人进行视频。而另一些人认为,社交软件的普及,会让人们的隐私权收到侵害,互联网公司会收集人脸信息、聊天记录等进行不正当活动。从功利主义的角度来看,社交软件的普及确实提高了社交效率与社交质量。但是反观大数据技术的普及,我们的隐私权受到了侵犯,这种潜在的威胁是否合理呢?因此用功利论来解决人工智能伦理问题是行不通的,我们必须引入系统功利主义。

4 人工智能伦理中的系统功利主义

系统功利主义是张华夏在《现代科学与伦理世界——道德哲学的探索与反思》中提出的。系统功利主义是为了调和道义论和功利论的矛盾而诞生的。系统功利主义既注重道义论中的善与义务,也注重结果是否符合功利原则,即最大程度上满足绝大多数人的幸福。^[15]关于如何量化系统功利主义,张华夏创造了一套函数。在计算行为功利时,用 x 表示行为本身,用 $U_d(x)$ 表示行为 x 的效用函数。计算准则功利时,用 x 表示行为本身,用 $U_r(x)$ 表示行为 x 的效用函数。将行为 x 的总效用函数表示为 $U_c(x) = f[U_d(x), U_r(x)]$ 。由于道德准则的不同所导致的道德行为的不同,张华夏用 R_1 、 R_2 、 R_3 、 R_4 分别表示功利原则、环保原则、仁爱原则、正义原则。 x 的效用函数表示为 $U_c(x) = f[U_d(x), U_r(x)] = DU_d(x) + RU_r(x)$ 。“这里 R 作为准则的功利系数, D 为行为的功利系数。 $R/D=k$ 为准则功利对行为直接功利的权重, k 为义理系数。 R 和 D 的数值是因行为主体而变化, K 的值反映了这两个系数之间权衡的重要差别。对于一个更注重准则功利主义的人, $k > 1$,而对于行为功利者来说, $k < 1$ 。”^[16]

陈晓平认为张华夏提出的系统功利主义思想十分有意义,但是只是能解决私德问题,但是系统功利主义的本质是调和功利主义和道义论之间的矛盾,功利主义强调的是大多数人的利益,这就不得不触及到公德部分。陈晓平的公式最初是引入盛庆珠的统合效应主义道德公式,模型表示为 $V_m(x) = K[V_o(x) + V_p(x)]$,公式中的 $V_o(x)$ 和 $V_p(x)$ 是客观价值, K 是道德系数。该模型的重点是引入了道德满足感 $V_m(x)$,由于道德满足感是因人而异的,所以道德系数 K 是随着道德满足感变化的。^[17]将统合效应主义融入到系统功利主义中,最终形成了分别用于计算公德和私德的道德函数模型: $V_s(x) = V_r(x) + V_a(x)$ 、 $U_s(x) = RV_r(x) + AV_a(x)$,需要注意的是,在实际应用中,需要优先考虑 $V_s(x)$ 函数,因为系统功利主义是基于系统论的一种理论,公德比私德有优先考虑权。

修改后的系统功利主义明确了公德与私德的问题,从道义论的角度来看,契合行为准则是“善”的原则;从功利主义的角度来看,发扬了功利主义的“满足绝大多数人幸福”的原则。系统功利主义很好地调和了两者的矛盾,避免了人工智能在做出伦理选择时,因为底层道德逻辑的混乱造成无法选择的局面。下面以火车改道例子来说明:假设一辆无法停止的火车在行使,一条路上有三个玩耍的小孩子,一条路上有一个铁路工人,假设不改道就会压死三个小孩,改道就会压死铁路工人,作为搬到员该如何抉择?如果多压死两个人,那行为功利为负数。若变道,则准则功利为负数,即少死两个人而破坏工作规则。根本问题在于压死两个年幼的孩子和触犯规章制度,到底哪个更具危害性,才是问题的根本所在?从系统功利主义的角度来看,两者的重要性是不相上下的。这个问题从公德的角度

来看就是无解的,就变成了私德问题,所以搬道员无论怎么做只要符合自己的心理上达到平衡就行,搬道员在道德上的自组织过程决定了事件走向,自组织过程也是内心感情因素最纠结的过程,比如搬道员对小孩有着怜悯的感情还是对规章制度的严格遵守。^[18]无论搬道员的选择如何,都不应该对其进行谴责。

系统功利主义是一种基于系统论的道德观点,它调和了功利论和道义论之间的矛盾,将私德与公德问题统一在一起进行解答。系统功利主义的特点就是避免了功利主义不为“善”而为“利”的局限,也避免了道义论一心向“善”的形式主义。世界上没有绝对的道德标准,人工智能在面临抉择时候,只能选择当下最适合大多数人道德准则的行为。目前来看,将系统功利主义作为人工智能道德底层逻辑是最合适的。

5 结束语

综上,以系统功利主义为主,义务论、道义论为辅,我们有希望整合出一条综合性的解决路径来发挥它们的优点、克服各自的缺点。采用系统功利主义的伦理准则,比如借助行动逻辑来约束智能体可选择的行动空间,从而最大限度地降低伦理风险。

参考文献

- [1] 陈晋.人工智能技术发展的伦理困境研究[D].吉林大学,2016.
- [2] 汪玉兰,邓国民.国际人工智能伦理研究现状与发展趋势[J].图书馆,2020(01):44-52.
- [3] BonnefonJF, ShariffA, RahwanI.TheSocialDilemmaof AutonomousVehicles[J].Science, 2016, 352(6293): 1573-1576.
- [4] O'NeilCathy.WeaponsofMathDestruction:HowBigDataIncreasesInequalityandThreatensDemocracy[J].SCIENTIFICAMERICAN, 2016, 315(2).
- [5] AllenC, SmitI, WallachW.ArtificialMorality:Top-down,Bottom-up,andHybridapproaches[J].EthicsandInformationTechnology, 2005, 7(3): 149-155.
- [6] 周欣卉.人工智能及其哲学思考[D].南京理工大学,2019.
- [7] 王东浩.机器人伦理问题探赜[J].未来与发展, 2013, 36(05):18-21.
- [8] 段伟文.机器人伦理的进路及其内涵[J].科学与社会, 2015, 5(02):35-45+54.
- [9] 周玄,赵建超.人工智能的伦理困境与正向规约[J].江西社会科学, 2022, 42(10):37-43.
- [10] 贺跃.康德道义论的目的论研究[D].湖南师范大学,2016.
- [11] 张曦.理解道义论:“牺牲者中心”进路与人的道德地位[J].道德与文明, 2014(01):41-48.
- [12] 王荣余.在“功利”与“道义”之间:中国人工智能立法的科学性探析[J].西南交通大学学报(社会科学版), 2022, 23(02):35-46.
- [13] 贾桢.人工智能发展的伦理审视[D].北京邮电大学,2019.
- [14] 程文俊.人工智能技术伦理问题研究[D].中原工学院,2020.
- [15] 邓环.系统功利主义的论证及其理论意义[J].苏州科技学院学报(社会科学版), 2015, 32(03):11-17.
- [16] 张华夏.论道德推理的结构[J].中山大学学报论丛, 2000, (02):142-154.
- [17] 盛庆珠.功利主义新论——统合效用主义理论及其在公平分配上的应用[M].上海:上海交通大学出版社,1996.
- [18] 陈晓平.系统功利主义及其应用[J].伦理学研究, 2012(05):17-23.