

【本期关注:人机对齐】

“镜子”与“他者”:智能机器与人类关系之考辨

彭 兰

【摘 要】在人工智能技术发展中,智能机器作为人类“镜子”与“他者”的角色正在不断凸显。作为“镜子”,智能机器帮助人类内窥到深层的认知规律,也在一些侧面反射着现实世界,映照出人类的局限。对个体而言,人机交流可以在一定程度上反射出“镜中我”,但这种反射有可能失真。机器也会强化、中介个体的自我凝视、他者凝视。从他者角度看,虽然早期智能机器往往被视作“低人一等”的“他者”,但今天人类需要更多地正视机器他者相对于人的对立性、差异性及其对人类的建构与掌控作用,进而促使机器成为负责任的他者。

【关键词】人工智能;智能机器;镜中我;机器他者;人机关系

【作者简介】彭兰,中国人民大学新闻与社会发展研究中心研究员,中国人民大学新闻学院教授(北京 100872)。

【原文出处】《新闻大学》(沪),2024.3.18~32

【基金项目】成果受中国人民大学2022年度“中央高校建设世界一流大学(学科)和特色发展引导专项资金、中央高校基本科研业务费”(2022XWTD001)资助。

ChatGPT、Sora等智能应用的先后推出,既引起一些人的欢呼,也触发了更多人对技术的恐慌。相比以往的任何技术,人工智能技术带来的不确定性都要更大,尤其在风险方面,但人类之所以走上人工智能的开发道路并且在这条道路上越走越远,或许并不完全出于莽撞。在人工智能这条新路上的探险,也是一种对人类本质与极限的探寻,在很多方面,智能机器成为了人类的“镜子”和“他者”,帮助人类加深了自我认识。这也是智能机器对人类的重要诱惑。

一、作为“镜子”的智能机器

智能机器在整体上是人类的镜子,通过它,人类可以窥视人类自身和认知的深层规律,同时也可以反观人类世界的某些面貌,以及人类的局限。对个体而言,智能机器也是镜子,这既体现在人机交流、互动中,也体现在它所中介、强化的凝视中。

(一)内窥深层认知规律的机器“镜子”

从技术角度看,人工智能的核心目标,是探求认

知规律,因此,人工智能被很多计算机专家视为“关于知识的科学”,核心是研究知识的表示、获取与应用(李德毅,2018:2),实质上是从机器模拟的视角来理解人类知识的相关规律。智能机器的发展,使人类对知识的生成、智能机器与人类思维(以及承载思维的大脑、身体)之间的关系,有了深入的探究。

目前人工智能有三大流派,符号主义(将现实世界看作是符号的计算,认为在符号计算上实现了相应功能,那么在现实世界也就实现了对应功能)、连接主义(它主要关注大脑的神经元机制,试图发现大脑的结构及其处理信息的机制,进而在机器上进行模拟)、行为主义(它的假设是智能取决于感知和行动,不需要知识、表示和推理,机器只要能生成相应行为就是具有了智能)(李德毅,2018:3-8)。这三大流派对智能、知识的理解都有自己的局限,但它们也具有特定的意义。三者从符号、大脑机制以及行为几种不同角度探究人类智能的本质、知识生成的路径,机器成为窥探人类认知规律的一面镜子。未来

的人工智能还会出现新的流派与技术,但机器这面镜子的作用肯定不会消失。

在理解知识生成规律的过程中,人工智能也迫使我们人类思维、情感、行为等与认知相关的“秘密”进行深入的“内窥”。尽管以往的生物学、心理学、哲学以及其他人文社会科学都在进行这样的研究,但人工智能技术让研究者跳出人类本身,从机器这样一个模拟者的视角来看人类社会的运行,以及人类的本质。新的视角催生新的认识,许多学科因为与人工智能技术的结合,也发生了重大变革,计算传播学、计算社会学以及计算法学等不仅是引入了计算的方法,更重要的是拓展了新的领域。

人工智能技术的演进也推动了“认知科学”这样的新的学科的发展。与人工智能三大流派非常类似,在认知研究中,也有三种典型的研究进路,分别是认知主义、联结主义和生成认知。认知主义认为认知是作为符号计算的信息加工活动;联结主义认为认知是由简单成分构成的神经网络整体状态的突现;生成认知则认为认知是一种生成活动,即一种主体与世界之间的结构耦合过程(孟伟,2020:35-36)。其中,生成认知在今天已经成为认知研究的主要进路。

人类在从智能机器角度认识自身的同时,也在试探着人与机器之间的最终边界,即机器对人类的学习、模仿可以到哪种程度,哪些方面是机器永远不能达到的。人们认为,那些机器不能达到的就应该是人类的本质,或者说“人性”。但在实践中,这个边界却在不断变化,人类对于自身的优越性、人类本质的自信也在不断被打击。

(二)反射现实世界的机器“镜子”

智能技术的应用,是为了解决现实世界的问题,这意味着机器需要尽力模拟人类社会和真实世界,很多时候,技术界会将现实模拟的真实程度作为技术水平高低的衡量标准。

对人类社会的模仿,必然以人类为学习对象。机器的语言表达,来源于人类的语料,机器表现出来的情感,是对人的情感的模拟,机器的行为模式,也

是对人类行为方式的模仿。机器的言行,也在总体反射着人类的价值观与文化。

机器对人类的学习,是基于具体的语料库或情境,面对不同的语料库或情境,机器获得的人类的素材可能大相径庭,有些语料库整体是文雅与文明的,有些则是充满戾气或污言秽语,由此形成的机器的语言与行为必然也会打上相似的烙印。

机器对人的学习并非一次性的,而是循环往复,人类的典型言行会经过机器的反复放大,产生“正反馈”效应。因此,“当我们在 ChatGPT 面前嘲笑其回答多么愚蠢的时候,其实真正愚蠢的是人类自己,因为在问答环节,ChatGPT 学习到的不仅是我们用文字提供的问题,也学习到了人类嘲笑它的态度和反应”(蓝江,2023)。

而从整体上看,机器在学习过程中,既可能复刻人类的共性,也可能放大人类的某些局部。例如,将一些枝节问题解读为关键问题,将个体行为推演为整体行为,将小众观念诠释成大众文化。机器是人类既有文化的放大镜与推广者,无论是对于美德的传承,还是对缺陷与错误的沿袭。机器放大与推广哪些局部,与机器后面掌握权力的人有关。因此,机器学习的结果,也反射着人类的权力结构。

除了通过学习来继承或揭示人类文化外,智能机器还可以用其他方式来反射现实世界。智能机器海量的信息生产能力,会将现实世界中发生的大大小小的各种活动、事件,变成媒介空间的信息,构建出“拟态环境”。尽管这仍然不可能完全复制现实世界,但相比传统媒体所构建的拟态环境,智能机器构建的拟态环境对人类活动的反映必然更为全面、丰富。大数据技术则可以通过社交媒体或数字空间的其他信息,来描绘社情民意的图景,解析社会发展的热点,或探知那些还在酝酿中的动向。

机器的反射不仅体现在人类活动方面,也在扩展到人类生存环境的各个方面。物联网将现实空间中的各类数据(如空气质量数据、人群密度数据、电器运行数据等)源源不断地提取出来,发送给需要这些数据的平台、机构或个体。“数字孪生”技术甚至可

以对城市等空间进行完整的复制。人们对生存空间的认知,也会更多地通过各种智能机器形成的数据及其加工产物。借助智能机器,现实环境也在“拟态化”。

Sora 带来的一个新话题是,机器能否完全自行模拟出现实世界。虽然由于 Sora 的技术原理没有公开,它是否已经是一个“现实世界的模拟器”人们还不能确定,但至少它让人们感受到了这种新的可能。从理论上来说,现实世界的万物运行既然可以用物理学规律来解释,那么用机器来进行模拟也是完全可能的,只是对算力是一个极大挑战,但这种挑战也并非不能克服。

如果这一技术实现,这就意味着,智能机器这面镜子中呈现的景象,不全是真实发生的,还会有很多是人造或“机造”的。智能机器的镜子属性,在未来也可能发生变化。甚至可能发生一种翻转,机器先制造了一种“镜中现实”,人类再去模拟它。

(三)映照人类局限的机器“镜子”

人类历史上曾经或正在使用的各种机器,都是对人类的某些弱点或局限的增强、弥补,每一种机器都是人类某些特征的反面映照,当然,在这些机器被开发出来之前,人类相应的特征或许并没有被当作缺陷看待,而技术的进步,使人类的不足与局限得到越来越多的暴露。智能机器同样如此。人工智能的发展,使人类不断反思自身在认知能力上的局限,并试图通过技术进行补充和增强。而在智能技术应用之前,人类对这些能力曾经是自信的。

从内容产业来看,智能技术的一个基本应用方向,是拓展人在信息采集方面的能力。信息采集维度与深度的拓展,也增强了人类对于社会现象和现实世界的认知能力。就像工业时代的机器一样,智能技术在信息生产、管理方面的效率也是它被广泛应用的重要原因,在信息爆炸的时代,这样的效率更是人类所需要的。Sora 因其具有“文生视频”的功能而备受关注,对于视频生产来说,这是一个关键性的效率提升。同样,ChatGPT 所解决的是基于海量信息的“量身定做”的信息整合与加工。即使它当下的水

平是平庸的,但已经能满足普通人的常规需求。在效率提升的同时,技术应用的门槛却不断降低,更多人可以参与到专业化的内容生产与创作中。

智能技术对人的认知能力的另一个补充,是对规律的深层挖掘、提炼,而前提是抽象问题的模型化与量化,即精确量化原本模糊含混的一些概念,例如特定领域中的“艺术风格”,博弈中的“棋风”,并且使这些只可意会、无法言传的技巧风格变得朴实无华,容易复制和推广(顾险峰,2016)。在将各种个性化的人类行为模型化、量化的基础上,这些行为的内在普遍规律便可得到一定程度的揭示。对人类行为的规律化总结,有利于相应行为的学习、模仿或大规模复制,同时也便于决策、调控,在内容生产方面也是如此。

从这些角度看,人工智能技术的出现也是信息时代发展的一个必然,它的开发目标,也是成为人类的补充者,而非人类的敌人。

但值得我们思考的问题是,在机器面前被不断揭示的人类的局限,是否一定都是问题,是否一定都要由机器去解决?人类是否一定要去追求某种能力的极限,获得这种极限能力是会让人类生活得更好,还是会更差?至少在今天,技术带来的能力增强,并没有让人们更轻松,更幸福。当我们获得了各种新技术增强的能力,例如强大的记忆能力(这建立在各种外部存储的支撑上)、随时随地的连接与交流能力、多线程任务的处理能力时,我们的压迫感与焦虑也在与日俱增。智能技术的未来发展,一定会减轻压迫感或焦虑吗,或许我们对此也没有把握。

智能机器这面镜子,不是在被动映射着人,它还带来一种可能,在不断增强人类的同时,造成人类对它的依赖,人类原来的短板会因此变得更短,甚至有可能接近“残疾”。它们在其他方面对人类的钳制也与日俱增,后文将从他者角度进一步讨论这一问题。但人类在一开始并不能完全预测这些风险,即使有所担忧,但总还是认为人类可以控制好智能技术,因为它们归根结底只是工具。然而智能技术比任何一种工具都具有能动性,它带来的不确定性也

是最大的。

(四)人机交流与“镜中我”

人工智能不仅是人类进行自我认识“镜子”,也是个体自我认知的“镜子”,这种认知主要通过人机交流、互动来实现。早在20世纪80年代,雪莉·特克尔便通过儿童与智能玩具互动的研究表明,展现有人类特征的人造物成为我们重新定义自我形象的镜子或“第二自我”(海勒,2019)。今天的智能机器不仅是玩具,也是人们日常交流的伙伴、助手甚至管家。对于个体的自我认知来说,它们作为“镜子”的作用会越来越突出。

米德指出,我们所称的“交流”的重要性在于,事实上它提供一种行为形式,使有机体或个体可以成为他自己的对象(米德,2016:157)。也就是说,交流使人们可以从他人那直面自己、认识自己。而库利则更明确地指出,“人们彼此都是一面镜子,映照对方”,人们通过他人看到一种“社会自我”,“这种社会自我可以被称为反射自我或镜中自我”(库利,2020:129)。

人要获得对自己的完整的认知,需要通过社会交往中的他者,但他者也是多种多样的,有些他者与自己存在较大的同质性,有些则具有较大的异质性。从同质性他者那里人们容易获得对自我的积极肯定,而从异质性的他者那里人们可以更清楚自己的优势与劣势,通过与异质他者的对比更好地定义自身的个性与特质,异质的他者也更容易激发人们的自我变革动力。但是,异质他者也可能带来对自我认同的负面影响,增加人们的认知失调,对此人们的本能是回避。

不同类型的社交互动,也会带来不同结果。今天的一些社交媒体,如微信,是强关系的社交,出于社会关系维护的需要,相互赞同、讨好则成为了互动的主旋律,但这并不会真正地让人们获得充分的自我认同,因为这个平台也充满了自我炫耀的社交表演,由此带来的社会比较压力与日俱增,这在无形中增加了人们的自我否定与内心纠结。而在另一些弱关系、匿名社交为主的平台,如微博等,人们之间的

相互否定、厮杀依然激烈甚至较以往有过之而不及,这也会增加人们的负面情绪。

社交越是频繁,人们对社交中的认同以及获得正面反馈的需要就越是强烈,这既包括在互动中来自他人的认同,也包括从社交环境的比较中获得的对自我的肯定与认同,但当下的社交平台并不能很好地满足人们的需求。

库利指出,自我最纠缠人的东西来自对他人意识的反映(库利,2020:173),保护社会自我的消极办法是逃避刺激和折磨人的暗示,积极的办法是与之斗争,学会控制和转化自我情感(库利,2020:175)。但人们往往更多地是采用消极的办法,当社交平台充满了刺激和折磨人的暗示时,逃避的动因也就不断增加,这会将一些人推向人机交流。与此同时,技术的发展,会使人机交流中的机器的“智商”“情商”不断提高。因此,可以预期,人机交流会逐渐普及。那么,在人机交流中,人对于机器会有什么样的期待,人们会在人机交流中对机器采取什么样的姿态?

有研究提出,用户认为被动而富有表现力的机器人更易于相处,而主动型机器人比被动型机器人更容易让人产生被支配感,甚至让人感到被不平等地对待(冯诀宵,2019)。一项针对儿童的人机交互实验也揭示,大多数参与实验的儿童都更喜欢与顺从型机器人而不是非顺从型机器人进行互动(Juan Martínez-Miranda et al.,2018)。为了迎合用户对人机交流的这样一种需要,在社交机器人的设计中,设计者的思路大多是以人为中心的,目前诸如小冰、ChatGPT、文心一言等也是以彬彬有礼、服从、迎合人的方式来与人对话。

在现实的人际交流中,人们常常不能完全选择自己的交流对象,也难以完全控制交流过程,但在人机交流中,选择性和可控性大大增强。

在人机交流中,人总是主动的一方,是诉说者和话题的引导者,而不会是倾听者。事实上在今天的人际交流中,人们对于他人的倾听也越来越稀缺,如韩炳哲所言,今天我们越来越丧失倾听的能力,妨碍倾听的罪魁祸首是日益严重的自我聚焦,即自恋。

将来或许会产生一种新的职业,叫倾听者,为他人奉献听觉并收取酬劳(韩炳哲,2021:107),但即使有人类倾听者这样一种职业,他们与“雇主”之间的交流中也难免会出现摩擦与冲突,这也给机器倾听者的大行其道提供了机会。

让机器充当倾听者,虽然对人类的诉说是有益的,但这种“主仆式”的交流模式却难以带来真实的“镜中我”。它帮助人建立起的自我认同,未必与人们在真实的人类交往中形成的自我认同一致,甚至可能因为机器总是迎合、取悦于人,这面镜子反射的自我会具有很大的迷惑性,这会妨碍人形成准确的自我认知,当人们回归现实的人际交流时,不恰当的自我认识可能带来交流的障碍。

库利指出,人在对待自己的社会自我的态度上有两种方式,一种是骄傲,一种是谦卑。骄傲是过于僵硬和过分自信的人肯定社会自我的方式,它隔绝了新的影响,使自己的成长受到阻碍,从长远来看,它会造成软弱(库利,2020:163)。可以想象,人机交流中,人们的骄傲情绪会比谦卑表现得更突出,但带来的长远影响,却可能是库利所说的“软弱”。

从技术角度看,人机交流中的机器完全可以被设计成具有异质性甚至挑战性的他者,但愿意接受机器挑战的人恐怕不会太多。即使人们希望机器具有一定的异质性与挑战性,也只是为了增加交流的乐趣,仍会以不动摇自我认同为前提。

(五)机器中介与强化的“凝视”

可穿戴设备等智能机器也会通过身体数据的采集与传播影响人的自我认识。身体数据不仅成为健康管理的依据,也成为自我传播的新纽带,成为映射“镜中我”的特别维度。

智能机器将人的自我传播扩展为物质自我(身体)与精神自我(意识)的对话。物质自我的各种数据,被精神自我获知,成为对自我状态进行认知的重要依据。而精神自我则会根据这些数据来做出行为调整或控制,如增加运动、改变生活方式或进行治疗等,这些手段也会对物质自我的数据产生进一步影响。

有了这些数据,人们仿佛可以跳出自己的身体,随时对自己的状态进行凝视。但这种凝视会带来对身体的过分关注,引发焦虑。原本微不足道的身体问题被放大,原来正常的的数据可能被人们不当的解读,机器误报的数据也会带来判断的差错。身体也不再是一个整体,而是变成一系列离散的数据。人们对身体的认知,不是来自自身的体验,而是来自数据性评估——是否达标、是否完美。对数据的控制、对完美数据的追求,替代了身体内在的真正感受。

对于个体而言,智能机器也具有拉康所说的主体的镜像认同中的他者的意义。拉康指出,主体的自我认同就像在一个镜像中看自己,这种观看并非主体自身的看,而是由他者的凝视结构出来的。在主体对镜像的观看中,不仅有属于想象界的自恋性认同,还有属于象征界的他者认同。后者是以他者的目光来看自己,按照他人指给自己的理想形象来看自己,以使自己成为令人满意的、值得爱的对象(吴琼,2010)。机器这样的他者,虽然不会像人类他者那样发表意见或表明态度,但它以数据方式对人进行的凝视,更为直接,更容易成为人的行为的“指挥棒”。

经过智能机器中介、控制的自我传播,也会延伸到人际传播中。人们将身体的数据(特别是运动数据)作为社会互动中自我表演的方式,并且与其他自我表演一样,期待从他人那获得积极的反馈,这时,机器也成为了人类他者对自我进行凝视的中介装置。

在现代社会,如吉登斯所言,对身体保持规训,是优秀的社会能动者固有的一种能力……对身体的习惯性控制,既是能动的一种内在本质,也被他人接受(信任)为个体的一种能力(吉登斯,2016:52-53)。在吉登斯那个时代,身体的规训更多通过外表、仪态等体现,而智能时代的身体规训则体现为运动健身、养生等,数据成为规训效果的主要证明。通过身体数据来进行身体的控制,成本可能更低(因为在一定程度上可以造假),且效果更直接有效。

机器中介或强化的凝视,看上去使人们增加了

对自我的关注与审视,但在这个过程中,人们受到机器和其他外在因素的影响也更加深重,人的自主判断与行动力则可能被不断削弱。

二、作为他者的智能机器

在作为“镜子”发生作用时,智能机器常常也会以相对于人类的他者的面貌出现。

在以往的哲学、社会学、传播学等学科中,有不少关于他者的理论,这些理论对他者的定义与阐释虽然有所差异,但也有一些共性。有学者总结,相对于主体,“他者”体现出三种不同的属性:其一为差异性,这种差异性体现为一种相对性;其二为建构性,主要表现为他者对主体的建构乃至掌控;其三为从属性,在某一对等级关系中,相对于高一级事物而言,低一级事物往往被认为是“他者”(胡亚敏、肖祥,2013)。今天,我们也可以参照这些属性去认识机器这个他者及其对人类的作用。但在对智能机器的认识过程中,多数人最早是把智能机器视作与其他机器一样的“低人一等”的从属者,当然这样的观点也不断受到机器的挑战,后来人们开始逐步关注机器相对人的对立性、差异性,智能技术发展到今天,人类才越来越深入地认识到机器对人带来的建构与掌控的力量。面对机器强大的建构力量,人类对机器又提出了新的要求,那就是需要机器成为负责任的他者。

(一)“低人一等”“从属于人”的机器他者?

尽管智能机器一开始就是以模仿甚至超越人类为目标,但到目前为止,多数人还是将机器视为比人更低能的对象来看待。人们也坚信,人的某些特质机器永远不能获得,人与机器之间总有一些鸿沟或边界不可逾越。

但是,这道边界究竟是什么,是思维、语言、情感、道德,还是身体及其体验、意向性、自我意识,抑或是创新能力?在智能技术发展的不同阶段,我们对此的回答是不一样的。而每一次智能技术的发展,都是对人类心目中的人机边界的突破,一些原本人类以为只有自己才具备的特质,也不断地被机器所模仿、复制。甚至连思考、语言这些能力现在机器

也初步具备了。

在《计算人类》一文中,凯瑟琳·海勒指出,美国学者福山坚信的人类在天地万物中占据独特地位的论点并没有足够的依据,她还参照人工智能的STA(sense-think-act,感知—思考—行动)范式,分析了一种可能的未来:人类与机器人难辨彼此(海勒,2019)。虽然我们现在还不能判断她预言的未来是否会真的实现,但机器越来越接近人,也是一个不争的事实。

对于越来越像人的机器,人类抱着复杂的情绪,“恐怖谷效应”或许是这些复杂情绪的一种体现:当机器人在外表、动作上与人类越来越像,人类会对机器人产生正面的情感;但若这种相似性到了某一特定程度,人类对机器人的反应会变得极为负面;而当机器人与人类的相似度继续上升,直至达到普通人之间的相似度时,人类对机器人的情感反应会再度回到正面(刘伟,2016)。“恐怖谷效应”是人类基于机器外表变化做出的反应,而对于机器在感知、思考、行动等方面越来越像人,甚至可能会变得跟人一致或更强这样一种前景,人们是否也会类似的态度波动?

当机器初步具备人的某些特征或能力时,人类会在一定程度上产生扩张与征服的快感。“当我们出让某些人格,将其赋予人工智能的时候,我们实际上是在积极地扩张人类自身,将人类从一种有局限的、受到限制的人类体扩张到其他非人类体上,这本身是形象拟人化所产生的独特功能,并且正是由于这一功能,我们从中得到极大满足”(王峰,2020)。当然,这种快感更多来自那些技术开发者。

但面对机器在能力上向人类的逼近,人类对机器的害怕、抵触、对抗情绪在不断上升。ChatGPT、Sora等应用的出现,便引发了普遍的恐慌,人们最主要的担忧在于自己的“饭碗”能否保住,以及人类与机器未来谁主宰谁。社交媒体也开始出现了类似这样的段子:“我们希望机器人帮人类扫地、洗碗,是因为人类要去写诗、画画。现在是AI都去写诗和画画了,我们人类还在扫地、洗碗。”一些人也感叹,人工智能的方向搞错了。

这些调侃当然是基于对人工智能的片面认识。作为“关于知识的科学”的人工智能,所要理解与运用的,不仅有写诗、作画这样的知识,也有关于体力劳动相关的知识。将人类从体力劳动中解放出来,也是包括机器人在内的人工智能发展的目标之一,扫地机器人等家务机器人也在不断进步。扫地机器人涉及的环境感知、人机交互等技术,同样是智能技术的体现。未来的人形机器人会帮助人们从事更多体力劳动。但人们对于写诗作画的机器的出现,比对做体力劳动的智能机器的出现,要敏感得多,即使今天写诗的机器的水平和扫地机器人的发展水平或许不相上下,但因为前者入侵了人类自认为只有自身才能涉足的领地,人们才会大惊失色。

在“人类中心主义”仍是主流价值观的今天,绝大多数人的愿望必然是,智能机器与其他机器一样,是服务于人的。但是,机器的进化是否真的如人们所期待的那样呢?机器发展走向的不确定性,给人类带来的恐惧正在与日俱增。

即使未来的智能机器可以按照人们期待的方向发展,将人们从各种体力劳动中解放出来,并辅助、激发人们去享受精神劳动的快乐,人们是否愿意完全接纳它们的存在?今天对此做出判断显然很难,但即使机器很完美,人类也不会放下忧虑,那时人们担忧的不是工作“饭碗”,而是人类的生存意义与价值。

另一方面,如凯瑟琳·海勒所指出,文化和技术的进化现在已经与生物进化如此融合,以至于它们不能再被视为孤立的过程。无论我们的未来如何,它几乎肯定会包括人类对生物过程的干预,这意味着“人性”将至少部分地取决于人类决定它应该是什么(海勒,2019)。也就是说,人与机器的边界,最终将由人类人为划定,并且通过人工干预使机器不能跨越这条边界,哪怕在技术上机器可以越界。但人类会如何划定这条边界,谁来代表人类划定这条边界,这也是一个充满不确定性的问题。

(二)与人类具有差异的机器他者

法国哲学家列维纳斯是关于他者研究的代表性

学者之一。他指出,形而上学意义上的欲望会引导着人类从自己所住的家向着一个陌生的他乡、向着一个彼处而去。这个欲望的终点是别处或他者(列维纳斯,2016:3-5)。这里的“形而上学”实际上指的是伦理学(叶秀山,2002),列维纳斯的他者理论的核心,也是他者伦理。

列维纳斯认为,在笛卡尔那里,我思是依赖于身为上帝的他者的,后者已经把无限观念置于(我的)灵魂之中,对灵魂进行教导(列维纳斯,2016:62)。而在社会交往中,“人的真正本质存在于面容中”(列维纳斯,2016:282),他所说的“面容”(也被译为“面貌”“面孔”等),是一种隐喻,象征了人的存在中一种非物化、非实体化的灵魂(张一兵,2004)。“面容”是他者出场的方式,是他者展现自己的方式(郭菁,2011),他者的“面容”是独特的,是我与他者的差别所在,也是他者的差异性所在(孙庆斌,2019)。社会关系也是在面对面中以从自我到他者的方式获得实现(列维纳斯,2016:281),这里的面对面,也就是“面容”对“面容”。“我们把这种由他人的出场所造成的对我的自发性的质疑,称为伦理。他人的陌异性——它向自我、向我的思想和我的占有的不可还原性——恰恰作为一种对我的自发性的质疑、作为伦理而实现出来”(列维纳斯,2016:14-15)。人要进行自我超越,就需要“通向他者”,以此将我从自身中解放出来(张荔君,2023)。他者不仅超越于我、差异于我,而且呼唤着我的回应。自我正是在回应他者的过程中,成为为他者的存在,主体性最终得以完全确立(郭菁,2011)伦理即意味着与他者的关系(许丽萍,2004:13)

列维纳斯将与他者的关系视为第一性的关系,强调他者伦理,是为了反抗传统的存在论哲学,在他看来传统哲学是一种总体论哲学,这种哲学充满了对他者的暴力,抹杀作为个体的每一个人的意义与价值,同化、还原并最终消灭总体之外的他者(列维纳斯,2016:318-319),而列维纳斯期待人们可以通过他者获得自我的超越。

即使列维纳斯等发出了解构一元总体论、拯救绝对他者的呼喊,但在今天的人类社会,如韩炳哲所

指出,一种普遍现象是“他者的消失”。他者的否定性已让位于同者的肯定性,由此造成同质化的扩散并对“社会体”造成伤害,而迁就和赞同也是伤害的原因(韩炳哲,2019:1)。当然,韩炳哲所说的他者与列维纳斯所说的他者含义未必完全一致,但他们至少有一种共识,那就是同一或同者意味着对差异、多样性的扼杀,而他者意味着多元性。

虽然在很多具体的情境中,如社会互动中,人们之间的相互否定、冲突仍然普遍存在,然而这些否定与冲突并不会改变人们的同一性,相互的冲突也是为了对抗他者带来的否定性,维护自我或所在群体的同一性,避免产生自我认同的危机。“数字化的全联网和全交际并未使人们更容易遇见他者。相反,它恰恰更便于人们从陌生者或他者身边经过,无视他们的存在,寻找到同者,志同道合者”(韩炳哲,2019:5)。当人们在社会活动中极力回避他者时,产生的是“充盈的空虚”,是荒淫(韩炳哲,2019:11)。

韩炳哲还认为,缺少他者的事物恰恰是被其对立、被不同于它本身的他者赋予生机和活力(韩炳哲,2019:9)。活力的产生,需要他者的“引诱”(韩炳哲,2019:11)。

人类活力逐步萎缩之时,智能机器的“引诱”也变得更突出,机器的效率和新的赋权对人们是一个提升活力的诱惑,机器认识世界的新视野也让人们感觉新奇。与此同时,至少到目前为止的智能技术,如ChatGPT,大都被设置为不具有客体(object)本身所具有的阻力或者对立、相对的感觉(刘海龙、连晓东,2023)。因此,它们最初是以一个非否定性的他者的面貌出现。即使一些人担忧这些技术对自己“饭碗”的影响,也会看到它们作为助手的价值,越来越多的人对这些技术是采用收编、驯化的方式,在这个过程中,人们也会获得某些新的活力。

但机器不具备对立性或许只是人们的一种错觉,这种错觉来源于开发者对于机器的“姿态”的设定。如前文所说,与人交流的机器表面总是顺从的。但从本质上来看,机器无疑会在某些方面成为人的差异者,甚至对立者。

智能机器虽然极力模仿人类的思维,但与人类固有的认知模式显然不同。智能技术的一切运作都建立在数据基础上,是数据化思维,这与人类的经验性思维,始终具有差异,某些时候也是对立的。人类既要接纳作为他者的机器思维,甚至利用机器思维来探索新的领域,但又不能完全被它所同化。对于智能机器与人的其他差异、对立性,人类的态度也应同样如此。

与此同时,我们还需要问的一个问题是,对于智能机器,人类究竟是要最终同化它,还是允许它保持并发展自己的他异性?今天的人工智能都是以对人的模仿为目标,但“成为人”是否就是它最终的归宿?如果人类开发智能机器的目标是最终把它们变成另一种人,那这种追求的意义何在?通过与人的差异性存在,使人的价值得到更多认识与体现,或许才是作为他者的机器更重要的意义所在。

(三)作为他者的机器对人类的建构或掌控

作为他者理论的另一位代表性思想家,黑格尔对于自我(self)或主体(subject)与他者(other,也译作“他物”)关系的观点是,自我或主体通过他者确立和认识自身。自我意识是从感性的和知觉的世界的存在反思而来的,并且,本质上是他的回归(黑格尔,1979:116)。

黑格尔对主体与他者的关系的基本思想,也往往被称为主奴辩证法。他指出,人有两种自我意识:“其一是独立的意识,它的本质是自为存在,另一为依赖的意识,他的本质是为对方而生活或为对方而存在。前者是主人,后者是奴隶。”(黑格尔,1979:127)在主人面前,奴隶感觉到自为存在只是外在的东西或者与自己不相干的东西;在恐惧中他感觉到自为存在只是潜在的,在陶冶事物的劳动中则自为存在成为他自己固有的了,他并且开始意识到他本身是自在自为地存在着的(黑格尔,1979:131)。尽管黑格尔所谈的主奴是两种不同的自我意识,但后来的研究者也将这样一种主奴辩证法延展到很多其他关系的分析中,这也包括自我与他者的关系。

黑格尔的思想启发人们,主人发号施令的行动,

可能在常规的看法中似乎是自由的体现,但实际上绝非真正的自由,反而是自由的障碍(杨云飞,2011)。尽管他者往往处于一种不平等的从属地位,但在一定条件下二者的地位会相互转化(胡亚敏、肖祥,2013)。奴隶主看上去能够强迫奴隶去干活迫使他服从自己的意愿;奴隶也必须放弃他的自我以满足奴隶主的要求;但奴隶可以通过他的劳动改变世界,同时也改变他自己而奴隶主却陷入了完全依赖于奴隶的境地(张剑,2011)。人与机器在很大程度上也是这样一种会发生转化的主奴关系。

对于机器、技术给人带来的控制,以往学者们已经有过大量的研究,每一代的技术发展促使学者的研究不断深入,而总体的判断,如列斐伏尔所说,“技术逐渐成为主人和舵手;如同资本主义制度下的货币和商品一样,技术呈现为一种自主的现实,思想、社会甚至国家控制不了技术”(列斐伏尔,2018:582)。也就是说,技术与人与人之间发生了主奴的翻转。

智能机器对人的控制,有些方面与传统的机器相似,但它也带来了新的控制逻辑与机制。

即使在发展早期,智能机器也并非总是被人类视为低等物种,在科幻影片中,人们常常想象人工智能作为超他者的发展、凝视与介入(黄鸣奋,2019),这种凝视与介入力量在现实中也在变得日益强大。今天,智能技术这样一种新的存在者虽然绝对服从人类,但是在实践中却将成为凝视人类的他者(刘海龙、连晓东,2023),机器的凝视既会强化人们的自我审查,也会成为外界对个体进行控制的中介。

前文已经分析了机器强化、中介的凝视对人类身体的控制作用。此外,人类的身体还会由于智能机器这样一个他者的中介,被其他更多因素所控制。例如,个体的身体状态数据成为平台进行商业推销的依据,劳动者的身体位置数据成为平台进行劳动控制的手段等,各种摄像头的“凝视”成为社会控制的方式。

智能机器的广泛应用,也带来了更深层的媒介化,生活、工作也成为媒介行为的一部分,人们的存在感、成就感也被一个至关重要的媒介数据所左右,

那就是流量。对流量的追求,不只是媒体、自媒体人的目标,也是许多普通人的目标。

智能技术也让人全息、全程数据化。包括身体、行为、情感甚至内在心理活动等在内的各个方面可以映射成多种维度的数据,这个映射过程会产生简化、筛选、聚光或消隐等效果,人的生命、生活的丰富“血肉”被抽离,变成抽象的数据“骨架”,在数据映射的基础上人还会被再抽象成有限的几个标签。通过标签,人与人、人与内容、人与服务等完成了聚类或匹配,原来人自身没有察觉的同质性、同类性被显性化。此外,数据还具有一种脱域机制,人的数据可以脱离本体,在外部进行重组,或被误读,或被越权使用。技术强化的数据化也逐渐渗透到人们的思维中,由此形成的数字拜物教也成为许多人思维的底色。

整体来看,人与社会现实被简化为数据符号系统,在此基础上,算法和其他智能技术影响着人们的生存、关系、认知,这些影响不断深化为对行为与思维的控制,更进一步,技术还会将控制变成监控——不是没有目标的泛泛控制,而在在监视基础上对每一个具体对象的精准控制。如韩炳哲在《非物》一书中指出的:“不是我们在使用智能手机,而是智能手机在使用我们。真正的能动者是智能手机。我们受到这台信息泄露设备的支配,在这台设备的背后有不同的能动者在驾驭着我们、控制着我们转向。”(韩炳哲,2023:39-40)

智能机器另一种潜在的控制,是将人们绑在不断加速的社会“战车”上。

智能机器的处理速度、效率,使人机交互中的延时越来越少,人们向机器发出指令,可以得到“秒级响应”,这样的节奏也会向社会互动扩散。但人的生理带宽不能像机器和通信带宽那样不断升级,因此,人们只能以全天候在线的方式来应对各种重要或不重要的任务请求,即使人们面对不同对象有不同的响应策略,有时也会有意进行延时处理,但总会有一些必须即时响应的任务。智能时代的信息、任务过载,也是人们必须做出快速响应的压力。尼采认为,

对刺激不即刻做出反应能力,属于“高雅的文化”(韩炳哲,2023:134),但今天的人们已经难以保持这种高雅。

德国学者罗萨在指出社会加速是一种新异化的同时,也分析了三个要素互动带来的加速循环:科技加速带来社会变迁的加速,社会变迁加速则推动生活步调的加速,而生活步调的加速又迫使技术进一步加速,如此循环往复(罗萨,2018:41)。在很大程度上,智能机器带来的效率,正是技术加速的突出表现,它在三个要素的加速循环中,起着尤为突出的作用。

尽管人们大多还保留着人类中心主义的优越感,认为技术是服务、服从于人的,但最后可能发现自己无力驾驭机器,反而会被机器掌控。“人类不是作为主体的目的和意志而存在,而是作为服务于技术物的能量和换能器而存在,人类的身体和精神与巨大的技术物的机制衔接在一起,构成了一台巨型机器(mega-machine)。”(蓝江,2023)

(四)承担责任的他者

作为他者的智能机器对人类的影响力量不断强大时,人类对它们也就不得不提出关于责任的要求。

相比其他机器,“智能机器可以承载使用者、设计者对其的期望,并且也被内置了感知世界并与世界互动的能力。在这个意义上,智能机器似乎有可能成为责任关系中的一方,并且承担某种程度的道德责任”(张正清、黄晓伟,2018)。在伦理、法律等领域,将机器视作责任主体或拟主体的观点也被越来越多地接受。

当然,智能机器也有技术上的差异性,在责任承担方面也必然有差异。研究者指出,人工智能应该包括有伦理影响的智能体、隐含的伦理智能体、明确的伦理智能体、完全的伦理智能体(段伟文,2017)。但无论机器本身能否成为道德或伦理的完全主体,但至少,人们期待人工智能的开发者可以将对人类负责的价值观和基本原则内嵌于智能系统中。价值对齐(value-alignment)便是人工智能技术界对这些期待做出的回应,即控制AI模型和系统使其符合人类

的意图、目标、偏好和道德准则(吴晓沅、谢幸,2023)。

但也有学者质疑,在现代性的境况下,“价值”本身是多元的、充满矛盾的、彼此冲突的,而非连贯性的、整体性的、系统性的。人类无法达成彼此之间(个体间、族群间、国家间、文明间……)的“价值对齐”,又如何使人工智能达成向“人类价值”的对齐?(吴冠军,2023)

赵汀阳则指出,人类的欲望、情感和价值观绝非好榜样,理性、情感、利益、信念何种优先,就难以排序,每个价值体系内部的优先排序也同样困难。假如让人工智能学会人类的情感、欲望和价值观,无非是让人工智能的意识陷于同样的混乱。给人工智能输入价值观只不过复制了人类的所有冲突(赵汀阳,2020)。

当我们试图让机器实现价值对齐时,我们却发现,最大挑战不在于技术,而在于人类自身。但即使面临极大的困难,如果人类不想被未来机器带入不可知的深渊,还是需要对机器的基本价值观做出必要的规制,当然,这前提是,人类在基本价值观方面取得更多共识。而作为机器开发者的技术人员,更需要为人类的未来发展进行自己的“价值对齐”。

三、结语

作为“镜子”的智能机器,映照出复杂的人与人类生态,面对这面镜子,人类可能在某些方面更清醒,在某些方面则更迷惑。机器也是具有极大能动性的“他者”,它们作为他者的差异性也是一面镜子,帮助人类审视自身。机器镜子本身也是凝视的他者,还中介着更多人类他者的凝视,这些凝视甚至可能发展成外在的控制。无论是人类整体,还是个体,既可能从机器以及它们所连接的更多他者那里获得能量与活力,也可能被这些外在力量进一步控制或同化,被人和机器共同构成的“巨机器”所吞没。最终究竟是哪个结果,取决于人如何利用机器的“镜子”和“他者”的积极意义,如何对抗其控制。

未来智能机器与人类的关系还会不断变化,是让机器向越来越类似人的方向发展,还是反其道行之?是坚持人类中心主义,还是拥抱人与机器平等

共存甚至人机共生的未来? 这些问题需要人类一次又一次的思考。但无论如何, 人类不应该失去对自身价值的坚守。

参考文献:

- [1] 安东尼·吉登斯(2016). 现代性与自我认同: 晚期现代中的自我与社会[M]. 夏璐, 译. 北京: 中国人民大学出版社.
- [2] 查尔斯·霍顿·库利(2020). 人类本性与社会秩序[M]. 包凡一, 王溪, 译. 北京: 华夏出版社.
- [3] 段伟文(2017). 控制的危机与人工智能的未来情境[J]. 探索与争鸣(10): 7-10.
- [4] 冯决宵(2019). 机器人表现力与主动性在交互中对用户的感知影响[J]. 成都理工大学学报(自然科学版)(1): 111-117.
- [5] 顾险峰(2016). 人工智能的历史回顾和发展现状[J]. 自然杂志(3): 157-166.
- [6] 郭菁(2011). 基于他者伦理的关怀[J]. 哲学动态(1): 65-71.
- [7] 哈特穆特·罗萨(2018). 新异化社会的诞生——社会加速批判理论大纲[M]. 郑作彧, 译. 上海: 上海人民出版社.
- [8] 韩炳哲(2019). 他者的消失[M]. 吴琼, 译. 北京: 中信出版集团.
- [9] 韩炳哲(2023). 非物[M]. 谢晓川, 译. 上海: 东方出版中心.
- [10] 黑格尔(1979). 精神现象学(上)[M]. 贺麟, 王玖兴, 译. 北京: 商务印书馆.
- [11] 胡亚敏, 肖祥(2013). “他者”的多副面孔[J]. 文艺理论研究(4): 166-172.
- [12] 黄鸣奋(2019). 超他者: 中国电影里的人工智能想象[J]. 江西师范大学学报(哲学社会科学版)(4): 65-76.
- [13] 凯瑟琳·海勒(2019). 计算人类[J]. 黄斐, 译. 全球传媒学刊(1): 28-46.
- [14] 蓝江(2023). 新唯物主义及其历史局限——从ChatGPT现象来看智能时代的物[J]. 甘肃社会科学(4): 18-26.
- [15] 蓝江(2023). 从技术治理到数字技术乌托邦——当代数字技术的控制机制和解放潜能[J]. 山东社会科学(11): 95-102.
- [16] 李德毅(2018). 人工智能导论[M]. 北京: 中国科学技术出版社.
- [17] 列斐伏尔(2018). 日常生活批判(第3卷)[M]. 叶齐茂, 等译. 北京: 社会科学文献出版社.
- [18] 刘海龙, 连晓东(2023). 新常人统治的来临: ChatGPT与传播研究[J]. 新闻记者(6): 11-20.
- [19] 刘伟(2016). 关于机器人若干重要现实问题的思考[J].

人民论坛·学术前沿(15): 35-43.

- [20] 孟伟(2020). 涉身与认知: 探索人类心智的新路径[M]. 北京: 中国科学技术出版社.
- [21] 乔治·米德(2016). 心灵·自我·社会[M]. 赵月瑟, 译. 上海: 上海译文出版社.
- [22] 孙庆斌(2009). 为“他者”与主体的责任: 列维纳斯“他者”理论的伦理诉求[J]. 江海学刊(4): 63-68.
- [23] 王峰(2020). 人工智能形象与成为“我们”的他者[J]. 上海大学学报(社会科学版)(4): 14-23.
- [24] 吴冠军(2023). 大语言模型的信任问题与资本逻辑[J]. 当代世界与社会主义(5): 4-14.
- [25] 吴琼(2010). 他者的凝视——拉康的“凝视”理论[J]. 文艺研究(4): 33-42.
- [26] 许丽萍(2004). 对列维纳斯他者伦理学的几点思考[A]. 中国现代外国哲学学会. 法国哲学的现代与未来全国学术讨论会论文集[C]. 北京: 人民出版社.
- [27] 杨云飞(2011). 《精神现象学》中的主奴关系解析[J]. 武汉大学学报(人文科学版)(4): 25-34.
- [28] 叶秀山(2002). 从康德到列维纳斯——兼论列维纳斯在欧洲哲学史上的意义[J]. 中国社会科学院研究生院学报(4): 28-35+109-110.
- [29] 伊曼纽尔·列维纳斯(2016). 总体与无限: 论外在性[M]. 朱刚, 译. 北京: 北京大学出版社.
- [30] 矣晓沅, 谢幸(2023). 大模型道德价值观对齐问题剖析[J]. 计算机研究与发展(9): 1926-1945.
- [31] 张剑(2011). 西方文论关键词他者[J]. 外国文学(1): 118-127+159-160.
- [32] 张荔君(2023). 他者的双重超越——列维纳斯思想中他异性在自我构成中的作用[J]. 哲学动态(12): 89-96.
- [33] 张一兵(2004). 大写他者的发生学逻辑[J]. 学海(4): 43-49.
- [34] 张正清, 黄晓伟(2018). 作为“他者”而承担道德责任的智能机器——一种人工智能成为道德能动者的可能性[J]. 道德与文明(4): 26-33.
- [35] 赵汀阳(2020). 人工智能提出了什么哲学问题? [J]. 文化纵横(1): 43-57.
- [36] Martínez-Miranda, J., Pérez-Espinosa, H., Espinosa-Curiel, I., Avila-George, H., & Rodríguez-Jacobo, J. (2018). Age-based Differences in Preferences and Affective Reactions Towards a Robot's Personality during Interaction. Computers in Human Behavior, 84, 245-257.