智能化背景下的算法信任

闫宏秀,宋胜男

(上海交通大学 科学史与科学文化研究院,上海 200240)

摘要:算法信任是技术信任的一种,隶属于传统意义上的人与技术之间的信任关系。算法与数据、智能的合力将这种技术信任推向了前所未有的境界。在数据被作为一种新型的生产要素的当下,作为数据核心动力之一的算法因其不透明性、偏见等技术不确定性问题所引发的算法风险构成了对算法信任的挑战。因此,需要从算法信任的缘起及其所蕴含风险的解析中找寻出路。有效的算法信任应当是算法自身的技术的鲁棒性与人类理性的契合。

关键词:算法;算法信任;调节;伦理;算法风险;算法设计

[**中图分类号**]TP391.3 [**文献标识码**]A [**文章编号**]1672-934X(2020)06-0001-09

DOI:10.16573/j.cnki.1672-934x.2020.06.001

Algorithm Trust under the Background of Intelligence

YAN Hong-xiu, *SONG Sheng-nan*

(*Institute of History and Culture of Science*, *Shanghai Jiao Tong University*, *Shanghai* 200240, *China*)

Abstract: Algorithmic trust is a kind of technical trusts, which belongs to the traditional trust relationship between people and technology. The join force from algorithm, data and intelligence promotes the technological trust to an unprecedented level. In case that data is regarded as a new type of productive factors, algorithm, one of the core driving forces of data, has led to algorithmic risks that pose a challenge to algorithmic trust due to such technical uncertainties as opacity, bias, etc. Therefore, it is necessary to find a way out from analyzing the causes of algorithmic trust and the risks involved. The effective algorithmic trust is supposed to be the combination between algorithm's own technical robustness and human rationality.

Key words: algorithm; algorithm trust; regulation; ethics; algorithmic risk; algorithmic design

在2020年4月出版发行的《中共中央国务院关于构建更加完善的要素市场化配置体制机制的意见》单行本中强调:“加快培育数据要素市场”,同时“充分体现技术、知识、管理、数据等要素的价值”[1]。可见,数据与土地、劳动力、资本、技术一样被视为生产要素类型的一种。数据、算法和算力三者的融合与整合是未来社会发展的必然趋势。

与人类对技术的信任一样,人类对算法的信任也相伴相随。因为信任是一种重要的社会整合力量,也是一个社会得以存续和发展的必要条件。在数据时代,算法信任是一种认知与

收稿日期:2020-08-16

基金项目:国家社会科学基金规划项目(19BZX043)

作者简介:闫宏秀(1974—),女,山西平遥人,教授,博士生导师,主要从事技术哲学研究;
宋胜男(1996—),女,浙江金华人,硕士研究生,研究方向为技术哲学研究。

实践相结合的探索。这种探索是人类寻求安全的一种方式。与此同时,在这种寻求安全的过程中,由算法信任所引发的问题而带来的信任危机在数据智能化时代越来越引起公众的广泛关注。那么,有效的算法信任何以可能呢?关于此,需要从问题的缘起着手,探索算法信任危机的根源及破解之道,寻求安全的确定性,构筑有效的算法信任。

一、算法信任的缘起

在人类历史发展过程中,信任与协作占据着极为重要的位置。在数据时代,信任依然是人类所面临的不可回避的重要议题之一。波兰社会学家彼得·什托姆普卡(Piotr Sztompak)针对"一些作者认为'人际信任'是信任的经典类型,通常在'社会信任'的标签下的其他类型的信任都只是派生的"[2](P55)这一观点,指出"现代技术,特别是电视,产生了对这种信任的一个有趣的种类——虚拟(virtual)个人的信任、程序信任、系统信任、对象化的信任(targeted trust)"[2](P56-P67)等。在开放的网络信息环境中,信任可被视为是"主观概率的一个特殊层级,一个代理依据此来对另一个代理或代理族群所将执行的特定行动予以评估"[3]。在智能化时代背景下,信任被引入计算机安全和云安全等相关领域,算法信任作为新的信任类型悄无声息地对我们生活的各个环节产生了重要的影响。伴随人类历史的发展,信任被不断地赋予新的内涵,技术信任就是对技术在人类社会发展中产生作用的一种表达。

(一)数据时代:算法信任缘起的语境

在人与技术共存的历程中,技术信任作为一个整体的概念出场。面对不同的技术,派生了不同的、具体类别性的技术信任。在技术对人类日益渗透的过程中,技术信任与人际信任一样成为了哲学反思的对象。

近年来,莫瑞奥萨瑞·塔迪欧(Mariarosaria Taddeo)与卢西亚诺·弗洛里迪(Luciano Floridi)提出了与彼得·什托姆普卡、尼克拉斯·卢曼(Niklas Luhmann)、拉塞尔·哈丁(Russell Hardin)等学者不同的信任范畴——"电子信任"(e-trust)。这种信任是在新技术背景下出现的一种信任类型,与传统人际信任最根本的区别是,其产生在"没有直接的、物理的接触的环境之中"[4]。此处的环境由数据科学与数据技术所构建。算法信任的出场,与"电子信任"如出一辙,都是发生在由数据科学与数据技术所构建的环境之中。

算法信任与电子信任一样,并没有像人际信任中委托者(人)与受托者(人)之间双方以在物理上的直接接触作为基础,都是行动者的交互,虽然并非都是直接的物理接触,但却真实存在,并可识别。算法信任与电子信任在数字化环境中的实现方式是相近的。一方面,随着各大互联网平台的兴起以及平台规范的制定,各大平台之间已经形成了一种互相可参考的普遍性互联网文化规范,在这样的情况下,算法信任无需以双方在物理上的直接接触作为基础。"一个能动者能够形成这样一种信念,这种信念允许其在未曾与受托者之间进行直接互动的情景下去相信一个能动者或客体(如自动服务)。"[4]另一方面,在数字化环境中,委托人甚至可以便捷快速地获得比通过信任双方直接接触更准确的信息。例如,我们可以根据官方邮箱的后缀来判定个人是否属于这一机构,可以通过各类失信查询平台查询他人有过的被执行失信信息,可以通过大数据搜索获得个人的相关新闻报道、社交联系方式以及可公开的知识成果。相比而言,在现实环境中,除非受托者主动呈现,委托者对受托者信息获取的便捷性及全面性远逊色于数字化环境。在数字化环境中,不论是作为人或者物的受托者都有自己的各类电子化证明,极少有能成功逃离数字化环境的记录和监督而存在的特例。

(二)算法信任缘起的两个维度

算法信任主要源自算法本身作为技术所蕴含的可信度和人们自身的信任倾向这两个维度。算法信任隶属于技术信任,其所涉及的对

象与技术信任一样，包括算法背后的人（算法技术的用户、设计者）和算法技术本身。

一般来说，信任来自委托者的给予与受托者的可信度。就算法信任而言，从委托者的视角来看，这往往有两种情况：

一是从算法技术的用户和设计者出发，算法信任以信任者的态度和期望为基础。在数字化环境中，算法技术的设计者和用户对于算法产品的熟悉度越高，这种信任就越容易产生，从本质上来说，这是基于“经验”的信任。但当算法处理复杂性问题越来越得心应手并展现出一定的可行性之后，信任者对于算法信任的基础由熟悉的“经验”转向对未来情况不确定的期望。

二是从信任本身的功能出发，“信任作为一种资本积累起来，它为更大范围的行为开放了更多的机会”[5](P85)。在卢曼那里，不管是人还是社会系统，赢得信任有利于处理和适应更加复杂的条件。同样地，作为数据时代特殊信任形式的算法信任也是如此，人们选择信任算法可以使得自身拥有更多的可能性。如帮助从事自身力所不能及的有关社会事件和决策等复杂计算，实现其完成某项任务的目的等。

从受托者的视角来看，即算法技术的角度来看，算法技术及算法产品的可信度是其获得信任的关键所在。虽然每一项技术都蕴含着潜在的风险，但技术发展本身恰恰是对风险的消解。因此，技术自身的鲁棒性催生着算法信任。与此同时，委托者与受托者之间的信任也是一个相互调节的过程。

（三）算法信任缘起的三个层次

从算法信任缘起的两个维度出发，又可分为必要性信任、期望性信任和算法可信任三个层次。三者关系密切，互相影响，任何一方面的失衡都会影响算法信任的规范性和安全性（如图 1 所示）。

第一，必要性信任。信任是社会系统运行的必要组成部分。“信任是系统成员之间相互作用的促进者，无论这些成员是人类代理、人工代理还是两者的组合（混合系统）。”[6]如果把正常的社会看作是一个运作系统，人们对医生、教师和司机的信任应该是自觉的并且也是必要的。这种必要的信任构成社会系统运行的关键。如果缺乏必要的信任，那么，授权就会遇到很多的阻碍，因为需要调配大量的时间和资源来对受托者进行监督。

因此，按照这个逻辑，任何执行任务的过程（包括监督的过程）也应该受到监督。然而，这就会影响大多数系统运行所负载的任务分配效率，限制系统整体的发展。在现实生活中，这种全方位的监督也是无法达到的。算法的应用涉及社会经济生活的方方面面。例如，匹配算法应用于网约车系统，推荐算法应用于信息个性化定制，决策算法应用于自动驾驶和机器控制等。信任是基于算法强大的工具性和计算性而建立的，同时，我们也需要信任来激发算法对于人类社会的巨大作用。但需要注意的是，这里的必要性信任不是意指对任何事物的无条件盲目信任，而是从系统运作的层面来看，必须有信任出场。

第二，期望性信任。信任是一个跨学科的概念，它“是一种心理状态，包含了接受脆弱性的意愿，这种脆弱性是指基于对他人意图或行为的积极预期”[7]。委托者的经验和背景以及受托者的能力是建立信任的重要方面。在这里，算法可被视为有能力成为受托者的角色。委托者对受托者的感知可信度是下意识的行为而非准确可靠的评估，委托者可以根据自身的直观感受来选择值得信任或不值得信任的受托者。也就是说，信任在某种意义上，可以被断定为一种主观性行为，甚至还有可能是无充分理性的主观性行为，即理性并非是期望性信任的充分必要条件。

当人们期望凭借算法给出的建议来进行公共决策和参与国家治理时，事实上就是一种算法信任。这种信任从本质上反映了人们对技术抱有的乐观态度。但与此相对应，也必然会出现技术悲观主义者，如“人工智能威胁论”“算法

陷阱”等消极观点。拉塞尔·哈丁认为,“信任是一种不道德的观念”,“潜在的信任者对信任的失败负有道德上的责任”[8]。实际情况是,信任往往会出现在不确定和危险的环境中,如果行动可以在完全确定或者零风险的状态下进行,那么此时的信任就没有任何意义。在算法时代,面对算法黑箱,我们需要保持谨慎并且乐观的期待,不断调节对于算法的信任来尽量规避不可预见的技术风险。

第三,算法可信度。必要性信任和期待性信任本质上都属于从人本身出发的以主观性倾向为主的信任,而算法可信度则是基于强大的技术背景、以客观性倾向为主的信任。“一个人在特定的场合信任一个特定的人,其原因并不在于信任自身的价值、重要性或必要性,而是在于这个特定的人的可信任度。”[9]从目前的算法应用来看,我们越来越多地把信任给予算法系统,然而我们并不理解其用于决策的方法,即所谓的算法黑箱普遍存在,输入与输出均已知,但输入如何被转换成输出却未知。在医疗判别、图像识别等领域,算法的可解释性与性能依旧无法兼得,这导致算法可信度更多地是出于算法的历史信用而非透明的可解释性技术验证。在人工智能领域,算法可信度体现在基于提高人工智能透明度和安全性的安全对抗测试技术和形式化验证技术的不断更新进程。

近年来,深度神经网络(DNN)成为人工智能应用领域的一个流行话题,继卷积深度神经网络(CNN)听觉模型之后,深度神经网络(DNN)成为一种更高效的判别模型,DNN算法将以往的识别率显著提高了一个档次。DNN可以理解为有很多隐藏层的神经网络(也被称为多层感知机,Multi-Layer Perceptron,MLP),一般来说,第一层是输入层,最后一层是输出层,而中间的层数都是隐藏层,它的高效率与高风险是并存的。针对伴随着算法高效率带来的风险,技术治理和伦理思考一起在不断地跟进,科技工作者展开算法可解释性、溯源、测试技术和形式化验证技术等方面的攻坚工作,人文社会科学工作者从伦理、法律、政策制定等方面进行算法伦理风险反思,出谋划策。算法的可信任度在备受关注中不断提高。

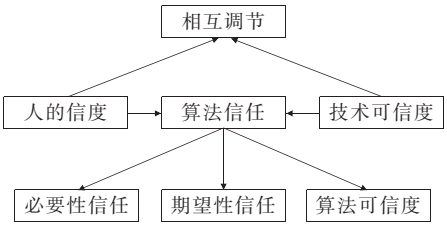

图1　算法信任缘起的两个维度和三个层次

以上两个方面的三个层次共同催生了算法信任。算法信任是一个动态的构成,即是一个持续性的过程而非最终的结果,在带来期待满足、效益提升等作用的背后,算法信任同时也存在着自身不可避免的缺陷,这种缺陷一部分源于算法技术带来的危机,另一部分则源自信任的情感基础和风险。

二、算法信任的风险及其后果

当算法在实际应用中出现重大决策失误时,“算法黑箱”和算法的不可解释性特征总是免不了被媒体和行业专家所聚焦。在现代计算机科学当中,“一个正确的算法是有用的,因为它可以解决给定的计算问题,一个不正确的算法也可以是有用的,与一般人认为的相反,不正确的算法在其错误率可以被控制的情况下甚至可能是很有用的”[10]。算法的这一特征在使用伪代码运行过程中可能会输出更好的结果,但是在实际场景的运用中却加剧了算法的复杂性和不可解释性。

然而,在市场竞争中,算法的不可解释性并不会削弱计算机算法软件市场的蓬勃发展。普通人已经无法忽视算法,算法正在大大影响人类的生活。实际上,任何技术都是一把双刃剑,但最终把握算法利弊走向的还是人类自身。“我们有可能把信任问题阐述为冒险,一种风险投资”[5](P31),算法信任同样如此。算法信任的

风险与算法风险相比更多地是透过算法技术直指信任、安全等伦理问题，算法风险更侧重于技术风险，其重在技术自身，而算法信任的风险则与委托者息息相关。

（一）算法信任程度过高导致盲目信任

在技术信任中，我们相信技术以及设计和操作技术的人。这种信任一旦过度，技术的权力将大大增加，因为过度的信任意味着委托者（信任者）将要求更少的证据和付出更少的监督。

当对算法的信任高达一定程度时，由信任初期带来的技术便利将走向另一端。我们通过数字技术与互联网世界中的人和事产生互动，共享信息，久而久之，技术会成为一个令我们足够信任的无形界面，在正常的日常生活中被我们使用却又被遗忘，只有在出现非人为事故的时候，我们才会想起平台、程序应用中最隐蔽的技术部分。当算法被嵌入关键系统，如自动驾驶、人脸识别、安防监控等日常安全设备当中时，我们常常会将其背后所蕴含的技术手段及其所存在的风险遗忘。

2016 年 1 月 20 日，一辆处在自动驾驶（Autopilot）状态的特斯拉在京港澳高速行驶过程中出现道路安全情况误判，引发车辆发生追尾导致人员伤亡的严重后果。这是全球首起特斯拉自动驾驶致死案[11]。智能技术的失控引起人们的反思，“算法黑箱”存在已久，之所以仍无法阻挡技术试跑，很大程度上归因于人们的信任。流血的错误将人们对自动驾驶技术的信任短暂回收，但这种不信任不会一直持续下去，因为科技公司绝不可能因此终止自动驾驶技术的研发，放弃该技术带来的商业红利。自动驾驶技术的目标设定本身是给驾驶员提供技术便利的驾驶辅助，但是由于特斯拉官方的不当宣传和驾驶者对于技术不恰当的信任，从而导致了事故的发生。“信任构成了复杂性简化的比较有效的形式”[5](P10)，但在数字时代无实体接触且仍不可解释的算法技术背景下，盲目的信任无异于随意掩埋威胁人类安全的技术地雷。这是算法信任最大的也是最不可预测的风险。

（二）算法信任程度过低阻碍算法发展进程

在智能化时代背景下，算法越来越具有高度的复杂性和自主性，以学习算法为例，“我们把经验数据提供给学习算法，它就能基于这些数据产生模型；在面对新的情况时，模型会给我们提供相应的判断”[12](P1)。学习算法本身不像传统的算法模型一样固化和简单，在发展过程中也越来越不偏重以人工先在设计为主，学习算法根据数据自主定制模型的自由程度也越来越高。

目前，学界对于算法的自主性和复杂性特征有着较大的争议。不同人群对于算法信任的程度也有所不同。一般而言，哲学、伦理学、法学、管理学等方面的专家会对算法信任更加谨慎，认为算法智能程度有可能冲击人类决策和人类福祉。正如阿里尔·扎拉奇（Ariel Ezrachi）和莫里斯·E. 斯图克（Maurice E. Stucke）指出：“在复杂的计算机算法、人工智能和大数据技术的辅助下，共谋、行为歧视与竞合场景将改变市场竞争范式并有可能恶化市场竞争环境。”[13]因此，他们呼吁政府和执法者发挥作用，投入资源以更好地帮助人们理解数据与算法所发挥的作用，增加执法部门对算法的管控；而计算机技术专家多对算法格外青睐，尽管出现 AlphaGo 等人工智能制胜的案例，但从整体上来说，算法目前还远不能及人类智能并且仍然是安全可控的。

如果把算法比喻成一个西瓜，那么人文社会科学学者更多地是从西瓜的生长土壤和培养环境以及它初显的皮相来判断西瓜是否甘甜。而计算机技术专家的判别逻辑则比较简单直接，采用“先尝一口”的方式，在技术可以实现的环境下先运行起来，允许算法有一定的容错率。哪怕存在若干复杂的不确定因素，也依然可以

试运行,因为算法应用对于算法技术的持续发展是至关重要的。“主张选择与经验观察一致的最简单假设”的奥卡姆剃刀原则在“机器学习领域也有很多追随者。但机器学习中什么是‘更简单的’这个问题一直困扰着研究者们,因此,奥卡姆剃刀在机器学习领域一直存在着争议。需要注意的是,奥卡姆剃刀并非科学研究中唯一可行的假设选择原则”[12](P17)。因此,算法本身的复杂性和不可解释性不应该导致算法信任清零。信任程度一旦变得过低,在人和技术之间的合作关系中,人就需要更多的介入。即信任程度越低,人工干预所需要的成本越大。

(三)不可拒绝的算法信任削弱人的主体性

鉴于算法信任的不可解释性,或许可以将其称之为不可拒绝的裹挟式信任。在某种意义上,算法信任恰恰是源于算法对人类的规训。算法拥有并且处理大量的数据,算法所触及到的视角比任何个人的视角都要广阔。米歇尔·福柯(Michel Foucault)在《规训与惩罚》一书中阐述了现代社会中的一种新的权力形式:规训权利。书中提到的“全景敞式主义”[14]不禁让人联想到数字化时代人的境遇,后台与用户有关的数据几乎全部暴露在算法之下。

迄今为止,算法在日常的观念和实践中不断规训人们养成新的信任习惯。当你下载使用某一个APP时,必须先接受其规定的技术设定和使用条款,否则将无权使用;当你搜索信息的时候,必须先输入符合搜索条件的关键词,否则将很难获取符合条件的有效信息;当你在网络上生产内容时,算法通过自动化指标对生产内容加以严格筛选和控制以引导用户产出符合平台预期的结果。出于商业需求、技术竞争和社会便利而被设计者训练出来的算法,在长时间发挥作用的过程中向人们传递了一个重要的信号,即权力信号。算法技术裹挟着规训与强制,在日常应用中以强植入性的方式来获取人们的信任。一方面人为的设计和制度的规定导致了这种强制性,更重要的是算法技术自身带有的长期以来的说服效应,它规训着人们对其养成习惯性的信任,俗称“口碑”或者“信誉”。思维习惯和行为习惯一旦养成,那么将很难改变,人的主体性也会逐渐丧失。

当算法越来越成为程序应用中最隐蔽的部分时,便产生了如下后果:首先,算法完全脱离了人们日常可感知的领域,进入到抽象意识的模糊领域;其次,算法的效力逐渐源于技术带来的必然性,而非可公开透明的算法模型和程序设计;最后,算法信任的确定性将算法效力发挥到极致从而阻止人们对算法本身产生质疑,进而当事故发生的时候,算法不会被推到第一责任方的位置。

因此,算法信任的风险是多种形式并且是难以预测的,但从信任的角度出发,算法风险往往集中在两个方面:一是算法信任的结果是不确定的,过分信任与缺乏信任都会导致技术发展过程中对自身要求的调整,从而产生不可预计的后果;二是算法信任遮蔽了算法自身的缺陷,即算法的复杂性和不可解释性等问题,给算法找了一个掩盖缺陷的借口。

信任的本质是允许不确定性的存在,算法信任也同理。伴随着技术研发的不断前进,一些不可解释的技术疑点会被信任搁置,带来潜在性的风险。不可否认的是,技术想要继续前进,必须要建立足够的信任,在计算机世界中,“数据”即为“经验”,对复杂的经验进行理性分析和寻找规律对于规避信任风险是必要的。

三、有效算法信任的构建路径

伴随着数字化进程的不断深入,算法已经成为构建未来世界的基础性原则。近几年,旨在加快发展人工通用智能的类脑计算系统发展很受重视。智能算法处理的行径与人类思维相似,呈现出生物性特征。在快速发展的需求和趋势下,算法技术不断实现突破,但在实际应用中却产生了歧视、偏见和黑箱等问题。这些问题在算法伦理研究中常常被提及,并且被视为

人工智能发展对人类产生威胁的一大隐患。

算法信任的偏颇会引发自动驾驶安全性、市场竞争公平性和个体发展的主体性等一系列重要的伦理问题，从而阻碍数字社会的健康发展。因此，科学技术界需要携手政府、企业等重新构建正确的算法信任，突破算法伦理困境，以便更好地服务于数字经济和人工智能的发展。由此，本文将基于调节理论，从信任的缘起与风险出发，结合算法特性、运行逻辑以及它的实现（技术、程序）与配置（参与人工设计的应用程序），尝试从如下四个方面构建有效的算法信任。

（一）设置适度的信任阈值

算法信任需要设置适度的阈值。适度的算法信任是算法安全发展的保障。适度的算法应该是“好”的、向善的，是以保护社会公共利益为前提的。如果信任仅仅从某个企业或者某个平台的角度出发，那么很容易因为利益的驱使而出现信任的背叛，从而导致一些不可避免的公平性问题和隐私问题，影响社会秩序。例如，作为虚拟助手的苹果的 Siri 等，目的是使人们的生活更加便利。这些助手是通过所收集数据的研析，做出对用户行为的预判或导引。但不容忽视的是，基于海量的数据分析和训练，虚拟助手会在其雇主超级平台的要求下，分析人们的消费能力和偏好之后进行精准营销，这实际上是进行了市场遮蔽，从而限制了用户的自由选择以及发现更多商品的可能。

现实中，人们往往要经过很长一段时间才能从平台这种由算法驱动的共谋场景中醒过神来。在计算机系统中，信任影响授权与合作，违背信任必定导致其收回。为了避免上述情况，及时地回收信任也是规避风险的有效途径。因此，算法信任需要设置一个适度的阈值并且定期评估信任的成效。鉴于算法拥有一般技术所不具有的超越物理极限的能力，如存储和计算能力，因此，人类要确保算法的风险可控性永远把握在人类手中，避免因失控而对人类道德和社会秩序产生巨大冲击。

（二）调节算法信任以规避风险

算法迭代更新快速，人们对于算法信任的反应普遍落后于算法技术演变的进度。因此，我们需要根据实际，从运行效果、训练难易、是否稳定等方面来评估给予算法的信任程度，在形成适度的信任阈值的基础上对算法信任进行动态调节。

为了公平对待算法技术，我们将其看作与人接近的道德主体。这一观点在彼得·保罗·维贝克（Peter-Paul Verbeek）那里有所体现，“技术不再调节人类行动与决策，而是与人类主体混合在一起，导致了有时候被成为‘赛博格’的混合实体”[15]。维贝克对技术物的道德维度解析，开启了对人和技术的关系进行规治的新路径。算法是数据时代的关键技术，当我们在某些场景中把算法看作是具有道德主体特质的人工物时，我们对待算法技术的态度将会更加审慎。“智能技术的发展将会使得人工道德主体界定的模糊性增加，进而造成其权利与权益的混乱。”[16]人与技术的关系需要根据实际情况不停地改变，技术的权利与责任认定也需要不断更新。

当今算法技术的发展恰恰正是人类在不确定性中积极寻求安全，以确保人与技术之间保持良好的关系。信任调节与“道德调节”相同，强调技术本身的自主性、意向性，利用调节的手段将信任内嵌于算法当中，以便算法性能与德性得到更好地发挥。算法信任调节以算法信任阈值为基础，不断将适度的信任植入到算法技术当中，推动算法在监督与规范下健康发展，规避算法带来的技术风险与伦理风险。

（三）完善算法设计者的责任机制

算法信任实际上是对设计和训练算法的人的一种信任。在复杂的架构和场景应用中，用户的选择和判断能力是有限的，算法能否进行正确的道德决策在很大程度上取决于算法设计者。当算法偏见问题产生的时候，我们需要对

其进行追责以避免算法风险的重复发生和扩大化。例如,在美国刑事司法COMPAS系统的应用中,“COMPAS是一种基于机器学习的软件,用于评估刑事被告再次犯罪的可能性,并已被证明提供对美国黑人有强烈偏见的预测”[6]。COMPAS系统的可信度受到了强烈的质疑,算法信任危机由此产生。

算法偏见的产生应该是可察的,“算法注入偏见主要有三个环节——数据集构建、目标制定与特征选取(工程师)、数据标注(标注者)。算法工程师和数据标注者是影响算法偏见产生的重要涉及者,数据集是偏见产生的土壤。事实上,几乎每一个机器学习算法背后的数据库都是含有偏见的”[17]。想要避免算法偏见获得算法信任,应该强调数据收集和标注时的公平性,及时消除数据本身存在的偏见,将公平一开始就嵌入算法设计的环节。

设计者应有权获悉算法用途、作出伦理判断、明确开发责任和后果,防止能力与责任的背离。如果算法尚未能够完全确保如自动驾驶决策的安全等,那么就应该在设计上明确强调提醒用户该算法产品的不确定性,以确保用户的知情权和选择权。因此,一方面,我们应该强调算法设计责任和伦理规则;另一方面,应该提醒用户对算法产品保持理智和有效的信任。因为算法在执行公平、隐私、透明度等社会规范时将以算法成效和企业利润作为直接的代价,进而无法确保企业在关乎利益的关键选择上会偏向于消费者。因此,需要完善算法设计者的责任机制来有效助推算法信任。

(四)算法信任应联动道德与法律

信任是带有个人偏好的,个人针对同样的算法技术在某些应用场景下选择信任,在其他应用场景下则选择不信任,个人的选择标准也不尽相同,正因为如此,我们才需要规范算法信任以发挥其正向作用。

算法信任在数据时代下涉及很多新的伦理问题,“信任不可能是无一例外、全然有效的行为准则”[5](P113)。算法信任作为一种推动力想要有效影响技术的研发和使用,就需要与法律和道德一起协同合作,共同应对算法和应用系统引发的危机。如果说道德在算法设计和应用中起到了先行者的作用,那么法律法规则对算法安全起到了最后的坚守作用。在法律规范方面,国内外政府、国际组织已经积极采取行动制定了与算法技术相关发展规范,例如,2019年4月,欧盟AI高级别专家组发布了《可信任人工智能的伦理框架》,提出了四项伦理准则:尊重人自主性、预防伤害、公平性和可解释性[18];2019年6月17日,国家新一代人工智能治理专业委员会发布了《新一代人工智能治理原则——发展负责任的人工智能》,强调了和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理等八条原则[19]。

算法是人工智能的灵魂,人工智能的伦理规范包括对算法技术的伦理规范。显然,道德与法律法规的联动规范是人类应对算法危机提出的切实有效的方案。同时,算法信任应该配合法律法规和道德规范一起对算法实施作用,成为解决算法危机的有效工具。有效的算法信任应该尽可能摒弃个人偏好,与法律道德实现联动配合,督促算法技术健康长足地发展。

有效的算法信任是在长期熟悉算法及其应用的过程当中建立的,算法因其透明性、偏见等技术不确定性问题所引发的算法风险构成了对算法信任的挑战,算法信任问题的解决有利于推动算法危机的解决。

综上所述,算法信任应该设置一个正确的阈值,根据实际情况进行信任的动态调节,落实与算法设计者和操作者相关的责任机制,尽快建立与算法技术发展相关法律规范,防止与算法技术发展相伴而行的巨大风险挑战。

四、结论

信任问题是确保社会正常运转的重要问题。在智能化社会背景下,这一问题的研究意

义更为突出，算法可信度以及我们对于算法的信任程度决定着我们在当下开放网络的生活中所承担的风险大小。信任需要有明确的边界，在边界内，算法对人类产生增强作用；超出边界，算法对人类的作用则变成操纵。

数据、算法、算力是数字经济的基础支撑，智能计算将会是未来技术发展的重心。随着工业互联网的快速发展，智能计算需求呈指数型增长，我们应该尽快建立对于算法以及人工智能技术的有效信任机制，缓解算法危机，防范技术风险，形成伦理规范，进而确保技术发展的方向是向善的并且符合大众的利益需求。

[参考文献]

[1] 中共中央国务院关于构建更加完善的要素市场化配置体制机制的意见[M]. 北京：人民出版社，2020.

[2] [波兰]彼得·什托姆普卡. 信任：一种社会学理论[M]. 程胜利，译. 北京：中华书局，2005.

[3] Gambetta D. Can We Trust Trust? [A]//Gambetta D (ed). Trust: Making and Breaking Cooperative Relations[M]. Oxford, Basil Blackwell, 1998: 217.

[4] Taddeo M. Defining Trust and E-Trust: From Old Theories to New Problems [J]. International Journal of Technology and Human Interaction, 2009(2): 23-25.

[5] [德]尼古拉斯·卢曼. 信任[M]. 翟铁鹏，李强，译. 上海：上海人民出版社，2005.

[6] Mariarosaria Taddeo. Trusting Digital Technologies Correctly[J]. Minds and Machines, 2017, 27(4): 565-568.

[7] Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, et al. Not So Different after All: A Cross-Discipline View of Trust [J]. The Academy of Management Review, 1998, 23(3): 393-404.

[8] Russell Hardin. Trustworthiness[J]. Ethics, 1996, 107(10): 26-42.

[9] Pamela Hieronymi. The Reasons of Trust[J]. Australasian Journal of Philosophy, 2008, 86(2): 213-236.

[10] 潘贵金，顾铁成，曾俭，等. 现代计算机常用的数据结构和算法[M]. 南京：南京大学出版社，1994：2.

[11] 观察者网. 邯郸特斯拉事故致死案：公司承认案发时处“自动驾驶”状态[EB/OL]. https://www.guancha.cn/society/2018_02_27_448303.shtml (2018-2-27)[2020-02-10].

[12] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.

[13] [英]阿里尔·扎拉奇，[美]莫里斯·E. 斯图克. 算法的陷阱——超级平台、算法垄断与场景欺骗[M]. 余潇，译. 北京：中信出版社，2018：305.

[14] [法]米歇尔·福柯. 规训与惩罚：监狱的诞生[M]. 刘北成，杨远婴，译. 北京：三联书店，2003：236.

[15] [荷]彼得·保罗·维贝克. 将技术道德化——理解与设计物的道德[M]. 闫宏秀，杨庆峰，译. 上海：上海交通大学出版社，2016：172.

[16] 李旭，苏东扬. 论人工智能的伦理风险表征[J]. 长沙理工大学学报(社会科学版)，2020(1)：13-17.

[17] 腾讯研究院. 算法偏见：看不见的“裁决者”[EB/OL]. https://mp.weixin.qq.com/s/4mFaDBzxxDSi_y76WQKwYw.[2019-12-19](2020-04-04).

[18] European Commission. Ethics Guidelines for Trust Worthy AI[EB/OL]. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines.

[19] 中华人民共和国中央人民政府. 发展负责任的人工智能：我国新一代人工智能治理原则发布[EB/OL]. http://www.gov.cn/xinwen/2019-06/17/content_5401006.htm, 2019-6-17.