人工智能算法的伦理维度

王天恩

(上海大学 社会科学学部,上海 200444)

摘要:在新一代人工智能发展过程中,智能算法的伦理问题逐渐呈现出新的发展态势,一个伦理维度正在以整体渗透的方式越来越清楚地出现在科学和哲学的结合处。人工智能伦理维度的形成,在更高层次涉及伦理问题研究范式的转换。深入到人工智能伦理的算法根源,可以从这一伦理维度更真切地看到人工智能算法的伦理属性。正是根据智能算法伦理属性的理解,人工智能伦理研究领域形成了越来越系统的基础性和操作性智能算法伦理原则,基于这些伦理原则制定相应的智能算法伦理规则,提出人工智能算法的伦理规制举措。本文试图通过国外人工智能算法伦理维度及其当代研究走向的系统研究,在智能算法层次展示大数据和人工智能发展基础上造世伦理研究的总体发展趋势。

关键词:人工智能;智能算法;伦理维度;算法伦理;伦理规制;造世伦理

中图分类号:B82-057 文献标志码:A DOI:10.3969/j.issn.1009-3699.2020.06.007

由于人工智能伦理不仅包括人工智能与人类的伦理关系,而且越来越涉及人工智能本身的伦理问题,更具整体性的探索必须深入到智能算法层次。在智能算法层次,伦理不再是涉及特定领域的问题,而是越来越呈现为一个贯穿整体的维度。在人工智能算法层次,伦理问题所涉及的已经不是与智能体的理解能力甚至核心机制相分离的单纯伦理领域,而是与人工智能算法包括核心机制及其进化甚至机器智能体的存在性基础一体化的伦理维度。

人工智能伦理维度的最深层次就是智能算法伦理。人工智能本身伦理问题的更深入讨论必须切入智能算法层次,从人工智能伦理的算法根源开始。

一、人工智能伦理的算法根源

人工智能发展所引发的伦理问题可以大致划分为两种基本类型:一是通用人工智能发展给人类带来的存在性风险引发的伦理问题;二是专用人工智能引发的技术性伦理问题。专用人工智能引发的伦理问题又可以分为两类:专用人工智能的算法设计和应用伦理。算法的伦理问题作为人工智能的核心机制领域不仅贯穿专用人工智能伦理和通用人工智能伦理,而且随着其发展而将伦理维度不断引向深入,随着其自主性的发育发展,算法发展到智能算法阶段。

由于具有从单纯的工具到逐渐出现自主行为的发展过程,作为以智能算法为核心的智能体,人工智能和开放环境之间的相互作用可能产生设计时不可能预料到的结果,这就会带来两方面的困难:一方面,即使程序员做对了每件事,人工智能的本域特定行为也可能是不可预测的;另一方面,系统安全性的验证成为更大的挑战,因为我们必须验证系统正在尝试做什么,而不是能够在所有操作语境中验证系统的安全行为[1]。随着智能算法自主性的发展,这种情况会越来越突出。对人工智能行为预测和验证的两方面困难越来越清楚地表明,必须深入到人工智能伦理的算法根源。

从新一代人工智能发展看,伦理的算法根源可以追溯到大数据处理。人工智能的伦理问题根源于从大数据处理到作为智能体自主行动整个过程所涉及的智能算法。在相应的大数据处理过程中,伦理问题主要出现在三个时间节点:当大数据移到一组选定的变量时、当选定的数据移到由算法或伦理研究产生的结果时以及当这些结果移到明确的行动计划时[2],而当这三个节点的控制由算法执行,算法设计便具有重要伦理意蕴。

智能算法对大数据的处理是人工智能伦理问题产生的起点,由此开始了智能算法伦理的发展进程。从智能算法伦理的发展可以看到伦理的不

收稿日期:2020-07-13

基金项目:国家社会科学基金重点项目(编号:17AZX003);教育部哲学社会科学研究重大课题攻关项目(编号:18JZD013).

作者简介:王天恩,上海大学社会科学学部教授,博士生导师,哲学博士,主要从事科学技术哲学研究.

同算法根源及其分布,从而有助于进入算法伦理的系统和更深层次把握。

在智能算法自主性越来越强和算法伦理层次越来越复杂这一发展形势下,最近一项关于算法伦理的综合性研究——基于伦理算法根源——提出了一个"算法伦理图谱"(ethics of algorithms map),为讨论智能算法伦理提供了一个概念框架,在此基础上可以更好诊断和应对智能算法的伦理挑战。从智能算法的层次看,这一概念框架正是主要关注(半)自主智能算法,涉及三种相互关联的情况:图谱将所涉及算法用于(a)将数据转化为给定结果的证据(从而结论),然后将这一结果用于(b)触发和激励一种(就其本身而言,或当与其他行动相结合时)在伦理上可能不是中性的行动,这项以复杂和(半)自主方式进行的工作使(c)算法驱动的行为效果的归责复杂化。也就是说,智能算法发展到一定阶段,可以具有自主性,因此能以自主或半自主的复杂方式触发和激励自主行动,从而产生伦理非中性而难以归责的复杂行为。由此,这个概念框架从根源层次展开了算法的整个谱系,构成了一个"算法伦理图谱",确定了产生于智能算法使用的五种伦理关切:三种"认知关切"(epistemic concerns)和两种"规范关切"(normative concerns)。在这五种伦理关切之外,再加上"可追踪性"(traceability),实际上是从算法根源划分人工智能伦理问题的六个方面。

在这个智能算法伦理图谱中,三种"认知关切"包括"非决定性证据"(inconclusive evidence)、"难以理解的证据"(inscrutable evidence)和"误导性证据"(misguided evidence)。"非决定性证据"源自更与符号主义进路相关的统计性;"难以理解的证据"主要源于解释机器学习算法使用的内在困难,这些困难存在于许多数据点中每一数据点所起作用上,包括导出结论、导致原则和实践限制;"误导性证据"则主要源于数据的关联误解,这种关联发生在数据和其不能说明的问题之间,数据共享给这种误导提供了条件。"规范关切"包括"不公正结果"(unfair outcomes)和"转化效应"(transformative effects)。"不公正结果"来自算法的伦理评价,"转化效应"所关注的则是智能算法结果的影响。算法可以影响我们如何使世界概念化,并改变其社会和政治组织。类如形塑过程的算法活动,通过以新的、意想不到的方式理解和概念化世界,并根据它所产生的直觉触发和激励行动重新本体化世界[3]。"转化效应"延伸到智能算法所导致伦理问题的后续效应,"可追踪性"则追溯到智能算法执行之前,这已经完全导向智能算法层次的伦理问题性质,六个方面构成了人工智能伦理问题的算法根源甚至是机制图谱。

深入到算法根源的伦理图谱,为更深入理解人工智能体发展的伦理维度提供了更直接的基础。在智能算法自主性发展及其相应伦理根源的基础上,可以大大深化关于人工智能伦理算法根源的理解。

关于人工智能伦理的算法根源,算法歧视的产生最为典型。通过实际案例分析,普林斯顿大学的弗里德曼(Batya Friedman)和尼森鲍姆(Helen Nissenbaum)对计算机系统的偏见及其根源作了系统研究,认为发展出了先前存在的偏见、技术的偏见和涌现的偏见三类:先前存在的偏见根源于社会制度、实践和态度,技术的偏见源于技术上的限制或考虑,涌现的偏见产生于使用的语境[4]。埃里克·戈德曼(Eric Goldman)的研究表明,偏见可能产生于在技术产生的"社会制度、实践和态度"中发现的预先存在的社会价值、技术限制以及使用语境涌现的方面。社会偏见可以被个别设计者有目的地嵌入到系统设计中,例如,可以在搜索引擎索引和排名标准的手动调整中看到[5]。如果做一个大致概括,算法歧视不仅源自设计者和使用者的偏见,而且源自作为根据的算法逻辑,因此算法歧视有两种基本类型:一是源自算法设计者成见的算法歧视;二是源自算法逻辑的算法歧视。

算法决策和数据权重设置导致的算法歧视一般都属于源自算法逻辑的算法歧视,这方面的实例在应用中越来越多。针对男同性恋约会应用程序 Grindr 的 GooglePlay 商店页面,被作为确定用户是否生活在性行为罪犯附近的类似应用程序推荐;Flickr 的自动照片标记暗示非裔美国人的脸标记为"猿";搜索非裔美国人的名字更有可能出现暗示此人有被逮捕记录的广告[6]。这些都属于无意产生的歧视,关于其根源,有研究做了系统总结:算法歧视可以有不同来源。首先,在算法决策中输入数据的权重可能很低,从而导致不同的影响。例如,作为一种间接歧视形式,预见性警务算法中过分强调邮政编码可能会导致低收入的非裔美国人社区与犯罪地区相关联,并因此而采用基于群体成员身份的特定目标。其次,使用算法本身的决定可以引发歧视。分类可以被视为直接歧视的一种形式,在这种情况下,算法被用于完全不同的处理。再次,算法可能因在不同情况下误

用某些模型而导致歧视。最后，在反馈循环形式中，有偏差的训练数据既可以作为算法使用的证据，也可以用来证明算法的有效性[7]。由此可以看到的更深层次内容则是：机器伦理的问题事实上与人工智能算法的发展、智能算法的进化机制问题“合二为一”。作为这一点的表现就是从专用人工智能到通用人工智能的发展问题，在这里，不再存在传统学科中单纯的伦理领域，而是只能看到一个整体存在的伦理维度，这一点，随着智能算法自主性的发展越来越明显。

作为伦理维度，算法的伦理问题随着人工智能的应用越来越普遍，而智能算法的根本伦理问题归根结底是由其自主性引发的。现代人工智能和机器人系统的特点是高度自治和不断提高的自主性，同时，它们在自动驾驶、服务机器人和数字个人助理等领域的应用也越来越接近人类。从这两种发展的结合中产生了“人工智能伦理学”研究领域，在这一领域，人们认识到自主机器的行为包含道德维度，并试图回答我们如何构建道德机器的问题[8]。道德机器概念的提出意味着智能机器与人类一样是伦理主体，进而其可以发展成为与人类伦理学一样的学科，正是由此，人们提出了人工智能伦理学。

人工智能伦理学可以理解为研究人工智能伦理问题的学科，也可以理解为人工智能研究的伦理维度。前者是关于人工智能的传统伦理学研究，而后者则属于科学和哲学一体化的伦理维度研究，这绝不仅仅是关于人工智能伦理学的狭义和广义理解，而是涉及科学和哲学甚至更广范围的重大范式转换，涉及两种层次不同的研究致思。随着智能算法的发展，伦理维度的研究越来越重要，这不仅是因为伦理研究关乎人工智能发展所导致的存在性危机及其人类应对，而且直接关系到通用人工智能的研究进路，这也是为什么随着人工智能的发展，相关伦理问题具有与以往完全不同性质的原因。正是在智能算法的发展本身，可以看到人工智能伦理问题的算法源起和发展，从而为深入系统研究智能算法的伦理属性奠定前提性基础。

二、人工智能算法的伦理属性

随着算法的智能化发展，人工智能算法的伦理属性将经历一个从人类投射到自身发育的过程。随着本身智能程度的提升，算法会被赋予一定的任务，从而表现出行为的目的性，这是人工智能算法的属人伦理属性。目前，公众讨论所关注的焦点是一类特定的决策算法，例如，在特定情况下采取最佳行动、对数据的最佳解释等，涉及数据和规则，增加或取代人类的分析和决策[3]。由此所涉及的主要还是人类伦理属性的人工智能映射，而自身伦理属性的获得使智能算法伦理进入了一个伦理发展的全新阶段。只有具有自身需要的智能算法，才可能具有自身的目的性；而只有具有自身目的性的智能算法，才可能具有类人伦理属性，因此人工智能算法的伦理属性由人工智能设计中体现人类需要和目的的算法赋予，并最终与智能算法的发展特别是自主进化密切相关。

早在人工智能研究之前，算法就已经经历了一个漫长的过程。“算法”一词源自公元 9 世纪撰写《印度计算法》（*Algoritmi de numero indorum*）一书的波斯数学家穆罕默德·穆萨·花拉子米，“花拉子米”（al－Khuwārizmī）一名就成了“算法”（algorithm）一词的来源[9]。随着人工智能的发展，算法发生了根本变化。在一开始，任何程序或决策过程，无论定义多么糟糕，都可以在媒体和公共话语中被称为“算法”[10]。算法可以是解决数学问题的运算步骤，甚至可以泛指解决问题的进程，但自从与智能相联系，就有了可以自己执行的智能算法。一旦具有自己执行的能力，智能算法的伦理属性便随即开始生成，而当智能算法开始自主进化并构成一个类群，机器智能就将逐渐获得自己的伦理属性——智能算法本身的伦理属性。

智能算法本身伦理属性的逐渐获得与人类构成了复杂的伦理关系，这属于一个新的研究领域——笔者称之为“智能伦理”，限于篇幅只能另文展开，本文主要讨论人为设计到自主进化智能算法的伦理属性。

在自主进化之前，智能算法所具有的只是属人伦理属性，这是人类伦理属性的人工智能映现，因此，算法设计会是越来越价值敏感的，算法设计领域甚至已经有了“价值敏感设计”（Value－Sensitive Design，VSD）的概念。

如果说，智能算法的价值负载开始都是无意识的，“价值敏感设计”则是有意识地纳入价值因素。价值敏感设计方法可作为一种工程教育工具（包括但不限于计算机工程），以弥合技术设计考虑与通过人类价值表达的伦理关切之间的差距。价值敏感设计以道德认识论为基础，通过一种考察概念、经验和技术问题的迭代三方设计方法，说明设计过程中的人类价值[11]。算法创构应当是价值敏感的，即对所创构的算法具有清醒的价值

意识。由于价值意识受着见识的极大影响,这又与道德想象力的未来把握密切相关。道德想象力是一种力量,它迫使我们将最高可能的现实和最大可想象的要求赋予一个不是我们自己的思想、行为或人,并且不以任何明显的方式接近我们[12]。对于道德想象力来说,越来越关键的是把握未来的能力,即未来把握力,因为一个行为是否符合道德,在越来越大程度上取决于对未来发展的预见,一种行为眼前看可能是道德的,但长远看却也许恰恰相反。如果说这一点在以往还不是那么引人注目,那么在智能算法设计中,伦理后果未来预见的重要性则达到了无以复加的程度,它以越来越大权重决定智能算法的伦理属性。

关于人工智能算法的伦理属性,可以通过算法歧视得到最具代表性的说明。目前所谓算法歧视,既有设计者无意导致,也有设计者或使用者有意为之;既可能产生于偏见,也可能产生于智能算法设计的合理性等因素。许多文献讨论到歧视是如何从有偏见的证据和决策中产生的,而一些社会偏见可能是无意的、更广泛的文化或组织价值观的微妙反映,例如,机器学习算法从人类标记的数据中训练,不经意地学会反映标记者的偏见。几乎在所有算法设计环节都可能导致无意歧视,分析可以无意中创建一个证据基础导致歧视[3]。分析环节都可能这样,由于具体条件抽离的根据和方式等,涉及概括的抽象描述可能更是如此。在智能算法中,无论用以训练的数据还是算法本身及其使用都可能造成偏差。一方面,算法组织系统中毫无疑问存在偏见;另一方面,算法反映了程序员和数据集的偏见。一方面,算法依赖的数据往往是有限的、不足的或不正确的[13];另一方面,偏差意味着不可靠,“垃圾进,垃圾出”(Garbage in—Garbage out)是对低质量数据输入导致数据输出不可靠的通俗说法[14]。这充分说明在偏差导致歧视等算法伦理方面,智能算法与大数据基础存在内在关联,因此,对歧视的关切已开始在关于大数据伦理的讨论中生根,这并不奇怪[15],而且,智能算法造成的偏差不是在数据造成的偏差基础上的简单叠加,在智能算法设计的可能偏差基础上,数据造成的偏差可能还会放大,产生不确定关系。人们日益认识到,用于促进各种操作的算法可以再现或造成偏差,这可能是因为算法的训练数据集本身在某种程度上有偏差,或者因为算法本身的操作产生了偏差。在涉及人工智能缺乏透明度的地方,这将是一个特别困难的问题[16]。由此可见大数据基础上智能算法的复杂性,从而看到这种复杂性基础上智能算法的伦理属性。

从算法的发展看,所谓“算法歧视”在不同阶段可以有不同性质。在专用人工智能算法中,由于歧视主要源自设计者的无意疏忽或有意偏见,因而只是源于设计者,而不是算法本身的歧视,确切地说属于算法产生的不公正现象,还不是真正意义上的歧视。只有当发展到通用人工智能,智能算法开始具有自主进化的能力,才会出现真正意义上的算法歧视。随着智能算法自主性的发展,会出现智能算法自发形成的歧视。

智能算法自发出现的歧视最初是从人类学来的。研究表明,人工智能可以从人类语言数据中习得偏见,因此在智能算法设计中,作为智能算法前提性基础,数据等条件的长远考虑具有越来越重要的地位,而其中最为根本的则是智能算法设计中的前提性预设,这与智能算法的人类理解密切相关。

算法偏见还与算法输出的人类理解和解释有关,算法的输出也需要解释(例如,一个人应该根据算法的要求做什么),对于行为数据,“客观的”相关性可以反映解释者的无意识的动机、特定的情感、深思熟虑的选择、社会经济决定、地理或人口影响。用这些术语中的任何一个来解释相关关系都需要额外应证——在统计模型中,意义并不是自明的。不同的度量标准使个体和群体的某些方面可见,而其他方面则感觉不到,因此,不能假定观察者的解释正确地反映了行为者的觉知,而不是解释者的偏见[3]。这是人类偏见在专用人工智能算法中起作用的方式,而在通用人工智能算法中,人类的作用方式则主要集中在最为关键的前提性规定环节。具体到算法歧视问题,就是必须有前提性规定层次的算法公正整体观照。

在智能算法层次,算法公正得到机制性凸显。算法歧视应当是算法公正的负面反映,可以归入更高层次的算法公正。智能算法的发展层次越高,越必须在更高层次考虑算法公正问题,由此所突出的,正是公正的层次性。智能算法的发展使算法公正越来越复杂,一方面,公正的层次性在智能算法中才得以真正呈现;另一方面,公正层次决定了公正的性质,低层次的公正可以是高层次的不公。有典型案例表明:值得注意的是,算法是否公正有时候与人们的直觉并不一致,比如在男女配对的配置算法中,如果规定只能男性向女性求婚,女性只能同意或者拒绝,那么结果是有利于男性。应该注意的是,这个算法有利于做提议的人

（在这种情况下是男人）[17]，这无疑是十分耐人寻味的，在婚配中，让男性主动女性被动符合大多数文化中尊重女性的观念，这看上去是公正的，但实际效果却恰恰相反，“尊重”导致选择的机会减少，选择的机会少导致来自“尊重”的不公平，这一案例典型地表现了智能算法公正的层次性及其所带来的复杂性，它意味着智能算法必须对数据等足够敏感，因此在智能算法中，广泛采取了一种称为“敏感数据”的视域，敏感数据不仅包括那些明确命名的变量，而且包括与它们相关的任何变量[15]。由于随着变量的增加和关系的复杂化，变量之间关系的把握会遇到“组合爆炸”，而“组合爆炸”意味着人类面临不可能把握所有具体关系的形势。智能算法设计之所以具有这么高度的价值敏感性，主要因为算法不仅非常敏感地涉及公正和歧视，而且涉及透明性和可解释性，特别是责任鸿沟问题，从而涉及人工智能算法的伦理原则。

三、人工智能算法的伦理原则

从长远看，在当代发展中，人工智能发展对人类的伦理挑战不仅最为严峻而且最为系统，系统研究人工智能的伦理原则，随着智能算法的发展越来越迫切。

在指导人工智能的开发和使用方面，在集大成基础上，微软最近系统定义了六项伦理原则，以确保人工智能系统公平、可靠、保护隐私、包容、透明和负责等。2018 年微软发表的《未来计算》一书，认为现在已经有可能开始定义六项伦理原则，以指导人工智能的开发和应用，这些原则应确保人工智能系统是公平、可靠和安全的，保护隐私和有保障的，包容、透明和可问责的[18]。《未来计算》所涉及的归根结底主要是智能算法，这是从智能算法发展长远着眼的必然结果，其所系统化的六项原则也是对相关领域研究具有代表性的概括产物。

在这一人工智能开发和应用的伦理原则系统中，可以大致看到三个层次：一是标准，包括公平和包容；二是基础，包括可靠、安全、隐私保护和保障；三是前提，包括透明和可问责。

在标准层次，“公平”处于六项原则之首。为了确保公平，需要通过培训，使人理解人工智能结果的含义和影响，弥补人工智能决策中的不足。公平问题源自人工智能具有某种程度的自主或设定选择的能力或功能，与前提性规定的设置有关。公平和包容密切相关，前提性预设同样涉及包容问题，包容所涉及的仍然是规定设置的合理性。为了确保可靠性和安全性，需要协调好人类过度信赖机器以及机器与人类利益之间冲突的矛盾。

基础层次的核心概念是可靠，没有可靠性，安全、隐私和保障都失去基础。在谈到可靠性时，伯朗格（Jérôme Béranger）提出了可靠性的概念不直接属于由新信息和通信技术产生的道德框架，但应由三个定义明确的参数提供说明：可由培训期间使用的数据通过其组成或资格产生的偏见、算法运行的有效性以及算法的几率性及其不确定性质和方面[2]。缺乏可靠性基础，安全性就谈不上，同样，隐私保护和保障必须建立在可靠性基础之上。可靠性和安全性等不是人工智能所特有的，只是越是层次高的自动系统，可靠性和安全性的问题越突出，而隐私等的保障则与智能化密切相关，因而透明性和可责性是人工智能伦理的前提性原则。在目前的专用人工智能研究中，可靠性和安全性都以透明性为前提条件。

前提性层次更为复杂，研究也相应更多。对于专用人工智能算法而言，以透明性为基本前提，因此在现阶段，透明性一直是人工智能伦理的重要原则，以至提出了“智能透明”（smart transparency）[3]的概念。在一般算法设计中，要求算法透明不仅重要而且具有合理性。算法不透明，现有关于人工智能算法的其他伦理原则就失去了前提。问题在于智能算法进入自主进化之后，自主进化的智能算法发展到一定程度，可解释性就会越来越丧失，从而透明性也就越来越不再可能。

智能算法的发展必定走向不透明，但关键的问题在于：在算法不可能透明的情况下，怎么做到可问责。在智能算法的伦理原则中，最能表明造世伦理性质的“可责性”不在造世伦理层次考虑问题，涉及人工智能的很多责任困境就不可能真正找到出路，特别是所谓“责任鸿沟”[19]困境，这也是因为人工智能的可责问题在更深层次涉及人工智能和人类作为一个整体的最重要关系。

机器学习开启的智能算法自主性发展进程带来了问责的严峻挑战。归责可能同时指向几个道德主体，这是在智能算法向自主进化过程中必定出现的问题，在人类智能设计的人工智能自主性不断增强的过程中，人工智能的应用必定经历一个责任主体的过渡衔接过程：由智能算法的人类设计者和使用者到完全自主进化的机器智能体。正是在这一过渡衔接阶段，必定出现非单一责任主体的现象，而多个道德主体势必导致复杂的责任关系，因此随着责任关系的复杂化，在智能算法发展过程中，责任鸿沟是一个必定要出现的重要

问题。

人工智能算法的透明性和可责性原则涉及更为复杂的内容,必须另文讨论。上述关于智能算法的伦理原则事实上都具有基础性质。除了这些基本原则,在智能算法伦理原则研究中还提出了一些操作性原则,这些伦理原则与人工智能算法的伦理规制更直接关联在一起。

提出人工智能算法操作性伦理原则的典型代表,当属伯朗格。在系统研究基础上,伯朗格提出了四项原则,包括"自治原则""仁爱原则""无恶意原则""正义原则",在这些操作性伦理原则中,又各自包含若干规则。伯朗格关于智能算法的四项原则,既是智能算法操作性伦理原则研究最具代表性的成果,也是这一领域的最新研究进展。

在四项原则的基础上,伯朗格建立起关于智能算法的十二个规则。伦理规则的目的是通过将命令强加于个人,在一个组织中建立秩序和社会结构,他们通过完善适当的伦理反思框架来防止无政府状态,它们的内容是可调整的,可根据语境、环境、空间和时间加以修改。通过以往的历史—社会—文化,每一个国家都有自己的伦理规则。通过给伦理原则以具体表达,使其存在于一定的情境和语境中,这些伦理规则包括一个令人印象深刻的结构内涵,它们是道德方法的调整变量,以应对特定的问题,特别是在数字世界[2]。自治原则包括两条主要的道德准则,即:尊重个人,将其视为充分参与决策或行动的行动者;自由意志,特别是与专业人员的专家自学制度结果有关及在不受外部影响和充分知情的情况下,自由和知情同意。仁爱原则与个别和(或)集体地为个人服务、为个人提供回报和额外利润、行为的效用和质量联系在一起。无恶意原则影响相称性的道德规则(因为我们总是在决定优劣或损益之间)、结果或行动在中长期的不确定性、预防和预期,特别是对一段时间行为的影响或后果的预期。正义原则与以社会团结而不是个人间互助的方式、行为者的责任以及行动的公平、平等和个人执行它的手段联系在一起[2]。智能算法的规制建立在人工智能算法伦理原则和规则基础之上,伯朗格基于其智能算法四项原则制定的十二个规则,为进入具体伦理规制奠定了系统基础,正是在这些原则及其所包含的规则基础上,伯朗格展开了智能算法规制的研究。

四、智能算法的伦理规制

探讨人工智能算法的伦理原则及其所包含的伦理规则,旨在智能算法的伦理规制。伯朗格关于人工智能算法的四项操作性伦理原则,为制定相应伦理规则提供了前提。

面对人工智能算法发展条件下新的伦理形势,伯朗格主要诉诸谨慎创新,他认为:这就是为什么必须在创新和谨慎之间找到一种平衡,这些伦理规则就说明了这一点,难道人们不能想象环境法模式下数字风险的责任原则?这一做法将使数字专业人员能够在合法数据、个人保护和更广泛的第三方数据权利范围内受益于这些使用。作为交换,如果他们没有采取必要的预防措施来避免这种风险,会使他们在受到重大损害的情况下承担责任,这将导致一个基于赋予行动者责任的双赢体系,此责任原则意味着算法或其用户考虑其过程和操作的影响[2]。创新和谨慎之间的平衡固然重要,但这还是表层考虑,当考虑到算法伦理所涉及领域之广泛时,问题就更为突出。由于主要源于创构活动,人工智能伦理问题的一个特点是涉及面广,伯朗格也注意到了这一点,这些算法系统的伦理规则包括所涉及的一系列领域,如个人数据保护和隐私保护、法律和监管方面、数据开放和共享、利益相关者、数据来源和算法分析应用以及规范和协议、技术发展、人机界面和接口、通信等。一种称为"公平"的算法应能考虑到其影响,并适当了解所涉及部门的特点[2]。事实上,智能算法不仅仅是单纯地涉及领域广,而是涉及领域的扩展表明了新的伦理特质,这种具有新特质的伦理问题,需要有新的伦理范式处理。

伯朗格的四项原则不仅基本上是传统伦理原则,而且主要是关于个体伦理的,这就决定了基于这些伦理原则和规则在为智能算法的伦理治理提供了理论基础的同时,又具有特定的局限。由于其传统伦理致思,受范式局限,伯朗格的算法伦理规制研究还主要在算法的传统伦理讨论范围,作为理论基础的四项伦理原则,正是其传统伦理致思的特定前提,但是,至少在算法伦理的定性探索中,伯朗格已经达到了研究的前沿,只是随着人工智能伦理问题的发展,算法根源的层次展开越来越清晰地表明研究范式转换的必要性。

随着人工智能越来越广泛的应用,在智能算法的伦理规制中,人们也意识到智能算法伦理的层次变化。由于智能算法在个人生活和社会领域越来越广泛深入的渗透,智能算法伦理规制的研究在层次把握的基础上,内部进路得到更好发展,越来越接近抓住问题的关键。

埃里克·鲍默(Eric PS Baumer)发现,智能

算法在日常生活和社会各个层面的普遍渗透使智能算法的伦理思考上升到了更高层次，他由此提出了一种以人为中心的算法设计方案，在基于算法的系统设计过程中引入了人和社会的解释。以人为中心的算法包括三个具体策略：理论上的、参与性的和思辨性的。理论上的策略（theoretical strategy）揭示两个重要方面：一方面，理论可以是规定性的，大多数数据集可能包含大量可能的特征；另一方面，理论可以是描述性的，给定一个算法产生的结果，它们意味着什么？理论可以帮助理清对这种系统结果的潜在解释。参与性的策略（participatory strategy）最主要的特点是“不会将可能受系统影响的人员作为设计过程的参与者”，而思辨性的策略（speculative strategy）则可以提供一种手段，用以克服我们目前对基于算法系统的看法[6]。从人工智能伦理算法根源的发展所展开的层次，可以看到只有到实现和配置阶段，算法才发展到智能算法阶段，算法伦理才发展到智能算法伦理，人工智能伦理的算法根源深藏在前提性预设之中，而在智能算法进入自主进化之后，对于作为最初设计者的人类来说，伦理规制就是关于前提性预设的尽可能长远考虑了——这就需要层次尽可能高的整体观照。

算法不仅具有典型双刃剑性质，而且越是涉及未来眼光越是如此。在伯朗格看来，算法自身具有相互矛盾的性质，它既是巨大希望来源（有用性、最优化、科学发现、速度等），又是重大社会关切（分类、保密和隐私、特征分析以及非人性化等）。因而在人类和智能算法的关系中，这是一个基本事实：关于建造和使用它们的方式，选择属于我们。这就是为什么在开发、实现和使用算法时，似乎有必要强加伦理规则，然后让开发人员和编程者们自己系统地考虑这些伦理规定[2]。当伦理规定的选择属于我们，我们有充分的选择权，这时候问题主要在当下的选择；而当智能算法进入自主进化，选择权就不再属于我们，这时候的关键则深化到了前提性预设。在这两种情况下，都涉及眼光远近，归根结底需要更高层次的整体观照，即使在算法歧视这种相对不算严重问题的具体伦理规制中，也是这样。

为了避免涉及性别、族裔和性向等“敏感属性”（sensitive attributes）与中性特征无意关联而触碰红线，罗梅（Andrea Romei）和鲁格吉里（Salvatore Ruggieri）提出了观察分析中防止歧视的四个重叠策略：①控制训练数据集的失真；②通过将其与反歧视标准相结合，修改分类学习算法；③提取分类模型后，对模型进行后处理；④在应用时，对预测进行更正，在受保护和不受保护的群体之间保持决策的相称性[20]。这是随着人类智能和机器智能在应用领域的融合而出现的特殊问题，为了让智能算法为人类谋福利而不是带来灾难，这些问题必须在智能算法的发展中得到实质性解决，这就进一步凸显了具体事物认识中整体观照的地位[21]。这一点，当智能算法设计前提性预设时更为根本，因为算法设计在根本上涉及前提性规定，包括价值和更基本的预设，这是智能算法伦理研究中的深层次内容。在这些方面，伯朗格做了很重要的工作。

在关于智能算法的伦理研究中，伯朗格概括出“为何、如何及多少”几个层次的问题，并在此基础上尝试建立起智能算法的伦理规则，他意识到，在建立关于算法的规则和伦理框架之前，似乎必须提供关于信息价值的概述和规格，信息价值具体取决于其使用的“为何、如何及多少”，该算法由具有固有基值的数据或信息提供。当它用于算法机制或处理的语境时，这个值将被更多地表达出来，并增加和获得它的全部潜力，这就把问题转移到超出技术划分范围的伦理层面[2]。将价值和规则相联系已经是对价值和规定的深入思考，不仅涉及价值负载而且包括最基本的前提性规定，这种前提性规定也不仅涉及文化，而且还会随着网络的发展在面上扩展。

在伯朗格那里，“为何”最为关键，它关系到目的，从而关系到需要及其发展，而对需要及其发展的把握则不仅涉及很多很复杂的因素，更重要的是涉及眼光，涉及对未来的把握，因此，“为何”必定导致对文化的质疑。算法依靠新的归纳逻辑形式，使决策和预测的经典模型受到质疑。有了算法，参与者不再局限于所属数据，而是向公共和开放数据库开放，特别是通过网络。的确，作为算法的前提性规定，不仅广及全球网络，甚至深及整个人类的社会潜意识。“如何”则涉及达到目的的手段，涉及技术但绝不纯属技术问题，因为如何达到目的往往涉及合法性和潜在风险等，至于“如何”，更确切地说是关于日常组织的问题。实施是使大数据的使用成为可能的基础，一个企业应该能够重新配置和发展，这取决于这些数字数据教给它什么，因此出现了一系列问题，诸如当异常数据流形成（监视）时，我们的 IT 安全系统如何快速通知我？我如何跟踪自己的数字数据？如何预测围绕我的数据生命周期而产生的负面影响、偏见和风险[2]？而“多少”则涉及隐私侵犯程度等。正是

涉及前提性规定的“为何、如何和多少”,使智能算法的伦理规制可能深入到更基础的层次,这与人工智能算法伦理规制研究发展的整体进路密切相关。由于涉及前提性规定,伯朗格关于人工智能算法的伦理规制主要属于内部进路。

根据近年来的研究,关于智能算法的伦理规制可以大致概括为内部进路和外部进路两种基本方式。

算法伦理问题解决的内部进路主要适用于解决工具性人工智能的算法伦理问题,在人工智能通用化发展层次,至少在某些阶段,问题的解决必须上升到更高层次。人们意识到:我们感到迫切需要采取行动,让来自不同领域的研究人员——包括法律、伦理学、政治哲学和机器学习——在现实世界中为不同任务设计、评估和验证其他公平标准[7]。与此相应,智能算法伦理问题解决的外部进路只适于各类人工智能的算法伦理问题的治理,其标志性特征是监管算法。

在某种意义上,监管算法进路遵循通过人工智能本身的发展解决人工智能问题的原则。最近有研究试图从可预测性入手解决问题,以此从“可预测性走向可验证的伦理机器人行为”。由于认识到确保自主系统合乎伦理地运作既复杂又困难,研究形成了由“后果引擎”组成的调控器,该引擎评估行动的可能未来结果,然后应用安全/伦理逻辑来选择行动。关于这一尝试,进一步的研究发现存在新的问题:拥有一个外加的“管理者”来评估系统所拥有的选项,并对其进行调整以选择最符合道德的选项这一想法,既很好理解也很吸引人,但不可能确定最合乎道德的选择是否真的被采纳,因而,在此基础上,他们又将一种著名的代理验证方法推广到其结果引擎,使之能验证其伦理决策的正确性[22]。这种外加“管理者”的做法,在专用人工智能发展阶段应当是管用的,但随着人工智能的通用化发展,“管理者”会渐失其作用和意义,也就是迟早会失效。外部进路只有在与内部进路构成双向循环的基础上,才能充分发挥应有的作用,人工智能的应用越广泛深入就越是如此。

关于人工智能伦理的算法根源、伦理属性以及从伦理原则到规制的研究取得了很大进展,只是人工智能发展所引起的越来越多伦理关注主要还是在传统伦理意义上的。智能算法层次的伦理问题必须更自觉地作为一个贯穿存在和过程整体的维度,在造世伦理[23]层次研究。从智能算法的伦理维度加以系统考察,可以更清楚地看到现有研究提供了重要思想资源,为提升到一个更高层次理解和解决新的伦理难题创造了条件。

在专用人工智能算法的伦理问题中,一些难题在人工智能和人类智能构成的更高层次视野中就可以看到新的出路,这一点在算法歧视问题的理解中最为明显。由于本身还不具有歧视观念,在专用人工智能中,还谈不上真正意义上的智能算法歧视,而只是算法公平性即算法公正在人类观念中的反映,因此,在专用人工智能发展水平的当下,算法歧视问题应当在考虑常规伦理规制的同时,更应该在造世伦理层次中加以考虑。在专用人工智能发展阶段,智能算法伦理问题的发生与到目前为止智能算法语境没有得到发育有关。由于人工智能语境发展的局限,越是在专用人工智能发展阶段,越可能发生类似歧视等伦理问题。只有人工智能语境得到相应发展,并与人类语境相融通,才可能在根本上有效避免这些伦理问题的发生。面对人工智能算法自主性的发展,人类必须尽早在更高层次关注和研究智能算法的前提性规定,不是在智能算法自主进化之后,而是现在,越早越主动。

不同于传统伦理问题,智能算法的伦理规制需要在新的伦理范式中进行。在人工智能算法的伦理维度,必须在人工智能算法设计等规律性探索过程中,通过逐步完善智能算法语境,并从个体伦理上升到类群伦理,才能真正面对智能算法伦理问题。正是在这个意义上,人工智能算法的伦理维度必须在造世伦理的层次把握,而这是关于智能算法伦理的全新研究领域。

在人工智能算法层次,伦理问题正在形成一个具有整体性的维度。由专用人工智能向通用人工智能发展,伦理维度越来越复杂,从主要是算法设计和使用者的伦理问题向超越人类伦理的人机伦理发展,从工具性人工智能的伦理规制走向自主进化人工智能存在性威胁的造世伦理应对,而在伦理维度的不断扩展中,则涉及越来越具有整体性的伦理问题。在专用人工智能发展中,透明性和可解释性一直是智能算法的重要伦理原则。由人工智能算法伦理原则推出的伦理规则是智能算法伦理规制的基础,不仅对于保持算法公正、避免算法偏见和歧视以及保护隐私,而且对于跨越人工智能应用的“责任鸿沟”至关重要。“责任鸿沟”的跨越关系到人工智能应用甚至人工智能本身的发展。在人工智能算法的伦理维度,个体伦理向类群伦理发展,传统伦理的界线正在“熔化”,这种变化趋势深及规则和规律的关系层次。在智

能算法深处，规则和规律越来越呈现一体化发展趋势。随着智能算法的发展，智能机器的道德主体地位及引发的相应伦理问题将构成对现有伦理观念越来越严峻的挑战，其应对必须建立在更高层次的伦理观念基础之上。人工智能算法伦理维度的发生和发展，具有深刻造世伦理意蕴，智能算法的造世伦理意蕴凸显了造世伦理的理论和实践一体化。规则和规律、理论和实践的一体化意味着智能算法的伦理规制越来越必须在造世伦理层次展开。

参 考 文 献

[1] Bostrom N, Yudkowsky E. The ethics of artificial intelligence[M]//Ramsey W, Frankish K. The Cambridge handbook of artificial intelligence. Cambridge: Cambridge University Press, 2014: 316－334.

[2] Béranger J. The algorithmic code of ethics: ethics at the bedside of the digital revolution[M]. London and Hoboken: ISTE Ltd and John Wiley & Sons Inc, 2018: 137－143.

[3] Mittelstadt B D, Allo P, Taddeo M, et al. The ethics of algorithms: mapping the debate[J]. Big Data & Society, 2016, 3(2): 1－21.

[4] Friedman B, Nissenbaum H. Bias in computer systems [J]. ACM Transactions on Information Systems, 1996, 14(3): 330－347.

[5] Goldman E. Search engine bias and the demise of search engine utopianism[J]. Yale Journal of Law and Technology, 2006, 8(1): 188－200.

[6] Baumer E P. Toward human－centered algorithm design[J]. Big Data & Society, 2017, 4(2): 1－12.

[7] Lepri B, Oliver N, Letouzé E F, et al. Fair, transparent, and accountable algorithmic decision－making processes: the premise, the proposed solutions, and the open challenges[J]. Philosophy & Technology, 2017, 31(3): 1－17.

[8] Berberich N, Diepold K. The virtuous machine－old ethics for new technology? [J]. Computer Science, 2018(6): 1－25.

[9] 瑟格·阿比特博，吉尔·多维克. 算法小时代：从数学到生活的历变[M]. 任轶，译. 北京：人民邮电出版社，2017：6.

[10] Hill R K. What an algorithm is[J]. Philosophy & Technology, 2016, 29(1): 35－59.

[11] Cummings M L. Integrating ethics in design through the value－sensitive design approach[J]. Science and Engineering Ethics, 2006(12): 701－715.

[12] Bromwich D. Moral imagination: essays[M]. Princeton: Princeton University Press, 2014: xi.

[13] Rainie L, Anderson J. Code－dependent: pros and cons of the algorithm age[R]. Washington, DC: Pew Research Center, 2017.

[14] Kilkenny M F, Robinson K M. Data quality: "Garbage in－Garbage out"[J]. Health Information Management Journal, 2018(3): 103－105.

[15] Goodman B, Flaxman S. European Union regulations on algorithmic decision－making and a "right to explanation"[J]. AI Magazine, 2017(3): 50－57.

[16] Boddington P. Towards a code of ethics for artificial intelligence[M]. New York: Springer International Publishing AG, 2017: 16.

[17] Dormehl L. The formula: how algorithms solve all our problems... and create more[M]. New York: Tarcher Perigee, 2014: 70.

[18] Microsoft Corporation. The future computed: artificial intelligence and its role in society[M]. Redmond: Independently published, 2018: 11.

[19] Matthias A. The responsibility gap: ascribing responsibility for the actions of learning automata[J]. Ethics and Information Technology, 2004(6): 175－183.

[20] Romei A, Ruggieri S. A multidisciplinary survey on discrimination analysis [J]. The Knowledge Engineering Review, 2014, 29(5): 582－638.

[21] 王天恩. 马克思的哲学革命及其内在逻辑的当代展开[J]. 江西师范大学学报：哲学社会科学版，2019(1)：49－57.

[22] Dennis L A, Fisher M, Winfield A F T. Towards verifiably ethical robot behavior[EB/OL]. [2019－10－28]. https://www.researchgate.net/publication/275055248.

[23] 王天恩. 大数据、人工智能和造世伦理[J]. 哲学分析，2019(5)：30－40.

[责任编辑　周　莉]