·科学技术与社会·

在伦理之前：重新理解机器学习算法的“技术缺陷”

Before Ethical Thoughts: Rethinking the “Technical Defects” of Machine Learning Algorithms

徐勐 /XU Meng

（武汉大学哲学学院，湖北武汉，430072）
(School of Philosophy, Wuhan University, Wuhan, Hubei, 430072)

摘　要：近年来，以机器学习为代表的算法在社会应用中引发了诸多伦理风险，由此产生了“算法伦理”这一问题域。它是从伦理视角出发批判算法技术的“非伦理性缺陷”，主张用伦理规制技术，解决道德难题。但通过梳理其技术原理及功能，本文认为“伦理优先性”批判对于技术而言是一种“苛责”或“强求”；质言之，算法导致的一些伦理风险是“先天技术缺陷”的必然结果，并非是由于缺乏伦理关照而人为造成的。这些“先天缺陷”主要有数据依赖、算法偏见和算法黑箱，它们应当首先被作为“前伦理”的技术问题来讨论，而非从伦理视角评判其是非对错。厘清这一关系有助于我们公允地讨论算法的伦理风险，并规避一些不切实际的建议，提出更具可行性的对策。

关键词：机器学习　算法　“先天技术缺陷”　伦理风险

Abstract: In recent years, machine learning algorithms have caused many ethical risks in the process of its social application, giving rise to issues of “algorithmic ethics”. It is a critical perspective from certain ethical standpoints of the “non-ethical defects” of algorithmic technology, and advocates using ethical principles to regulate technology in order to solve some moral problems caused by algorithms. However, by sorting out the operational logic and functional characteristics of machine learning, I argue that the criticism of “ethical priority” is an “harsh criticism” or “excessive demands” for algorithms. In other words, some ethical risks are caused by “innate technical defects”, not the consequence of lack of ethical norms deliberately. There are three main types of these “innate defects”, namely data dependence, algorithmic bias and algorithmic black box. They should first be discussed as “pre-ethical” technical issues, rather than judging their right or wrong from ethical perspectives. It is helpful for us to discuss algorithmic ethical risks within a reasonable range by clarifying this problem, and avoid some “unrealistic” ideas and propose more feasible solutions when giving countermeasures and suggestions.

Key Words: Machine learning; Algorithm; Innate technological deficiency; Ethical risks

中图分类号：B82; C01　DOI: 10.15994/j.1000–0763.2026.05.012　CSTR: 32281.14.jdn.2026.05.012

一、问题的提出：算法伦理研究中的概念框定及其现状分析

随着人工智能时代的来临，算法在社会中应用得愈发广泛，其重要性不言而喻；但同时也引发了一系列伦理问题，于是二者相结合形成了以“算法伦理”为中心的跨学科问题域。

收稿日期：2025 年 5 月 23 日
作者简介：徐　勐（1998–）男，江苏泰兴人，武汉大学哲学学院博士研究生，研究方向为科技伦理。Email: xm1219958815@163.com

从字词剖析来看，它是由“算法”和“伦理”组合而成，也就意味着两种研究进路，一是“技术–伦理”进路，从技术原理出发切入伦理后果；另一个则是“伦理–技术”进路，从伦理风险现状出发反思技术的非伦理性。针对现状，算法伦理研究有两个值得商榷之处，其一是对算法的技术性讨论略显不足，对具体的研究范围及其技术特征模糊其辞；其二则是重“伦理–技术”而轻“技术–伦理”，以伦理优先性为研究重心。

首先第一个问题，学界对算法伦理中“算法”的技术性讨论存在不足。尽管其内容在计算机或人工智能领域属于常识，但在伦理学界却鲜有重视，主要表现在对“算法”的界定模糊不清，对其原理的讨论也晦暗不明。在计算机领域，算法被看作用某种方法解决问题的策略机制，被具体化为一组准确且完整的描述或一系列清晰的指令。[1] 然而这仅是一种抽象定义，实际上其内部也有不同类别，比如机器学习（Machine Learning）与传统算法之间就有较大差异，哪怕是深度学习作为前者的子集来说，二者也有区分，因此笼统地用“算法”来一以概之不仅导致问题无法聚焦，而且也会出现文不对题、以偏概全的尴尬情形，因为并非所有的算法都会导致“算法伦理”讨论中出现的问题。而从一些论述来看，他们讨论的功能特征和伦理风险大多出自机器学习算法，部分源自深度学习，也有二者兼而有之的状况，比如著名的“算法黑箱”（Algorithmic Black Box），但这个问题却与传统算法毫无关联，后者显然没有这种顾虑。不过多数文章却对此模棱两可，让读者琢磨不透。不仅如此，或许是受制于学科差异，很多学者仅仅粗略地了解一些基本的社会应用，但却无法清晰地阐明算法的技术原理和运作逻辑，也就无法精准地把握其功能特征，这种“只知其表、不知其里”的讨论显然无益于进行更深层次的伦理研究。

于是，这一现状也造成了如今伦理学研究的窘境，即学界大多从“伦理–技术”进路切入讨论而对“技术–伦理”的模式把握有限。这点可以从他们对算法伦理的定义中看出端倪。通常来说，算法伦理是探讨算法如何在设计、开发和应用过程中保持与主流社会价值观相一致，确保算法应用不背离特定的社会伦理规范的研究。[2] 不难发现，“伦理–技术”模式往往起步于算法导致的伦理后果，从规范性角度批判这些现象，继而为了消除这些隐患，提出对算法进行伦理治理或社会规制。这种进路也是当下的主流趋势，其宗旨是“伦理优先”。比如有学者首先从社会现状出发，提出防控和治理机器学习伦理风险，需要遵循以人为本、公平公正、透明可释等原则，构建既高效又符合伦理道德标准的人工智能生态系统，确保技术创新与社会责任的平衡。[3] 也有学者关注具体的伦理问题，即算法不公正，然后为了更好地利用算法促进社会公正，需要深入剖析其根源，根据相关类型提出更具针对性的治理路径，引导算法朝着公正的方向发展。[4] 而从具体治理而言，有人主张在设计阶段就将伦理因素纳入考量，通过有效融合功利论、契约论、义务论和德性论等伦理思想，构建多元主体参与的治理网络，来实现政府、企业、公众等不同主体的协同治理。[5] 如此观点不一而足。

从合理性角度而言，这些想法未尝不可，但却造成了一种因果倒置的错位感。他们虽然关注到了伦理问题是由算法导致的，但却过犹不及，完全将重心放在了算法的伦理性之上，似乎忘却了它首先是一种技术工具。换言之，他们没有注意到一个事实，即导致伦理风险的某些技术性因素是“先天”的，并非是缺乏人为的伦理设计或治理而酿成的过错。比如我们在反思社会公平相关的算法歧视时，由于算法歧视是人们对于算法含有偏见的结果不经审查就作出决策的行为，是算法偏见的后果，[6] 所以为了避免算法歧视的发生，人们试图根治算法偏见。但实际上，算法偏见并不是我们可以用伦理手段来克服的，它是“前伦理”的存在，是算法的内在属性，是“先天的技术缺陷”。于是也就形成这样一个吊诡的现象：以伦理之名对算法进行的非伦理性批判成了一种不合时宜的“苛责”或“强求”。而通过分析发现，我们习以为常的数据依赖、算法偏见和算法黑

箱均属于此类“先天技术缺陷”，虽然它们极易导致隐私泄露、算法歧视、归责模糊等棘手的伦理问题，但这些缺陷本身在价值和规范上是中立的，用技术性手段来解决非是不可，实则不能，因此针对这些缺陷我们只能缓解或弱化其导致的伦理后果，而无法在技术上根治。

结合上述现象，本文试图挑战传统的“伦理–技术”进路，重新关注式微的“技术–伦理”进路，厘清机器学习算法的相关概念，这虽然是基础工作，却至关重要。所谓名正则言顺，言顺则事成，我们只有先理解算法的基本原理及其功能特征，才能去把握它造成的伦理后果，然后进一步提出更具可行性的治理路径。不过限于篇幅，本文只能完成前一半的工作。从结构上来说，本文首先将算法的研究框定在机器学习的范围[①]，详细梳理其技术原理和运作逻辑；其次讨论其功能特征及社会应用；最后指出，数据依赖、算法偏见和算法黑箱是算法的“先天技术缺陷”，它们并非是缺乏伦理设计或监管而造成的，而是本身的技术特性导致的，因此针对它们导致的伦理风险而试图从技术上根治的想法都是不切实际的。

二、机器学习算法的技术原理

从历史图景来看，机器学习是人工智能发展到一定阶段的产物。当下学界对机器学习的定义大多采用汤姆·米切尔（Tom M. Mitchell）的论述：假设用P表示计算机程序在某类任务T上的性能表现（即性能度量P，performance measure），若一个程序通过利用经验E在任务T中改善或提升了由性能度量P衡量的性能，则我们就说这个程序对E进行了学习。[7]这一定义较为抽象，更通俗地说，机器学习是致力于研究如何通过计算手段，利用经验来改善系统自身性能的学科。其中，“经验”通常以“数据”的形式存在，因此机器学习所研究的是关于在计算机上从数据中产生“模型”（模型指算法从数据中学到的结果）的算法，即“学习算法”（learning algorithm）。（[8]，p.1）当然，学得模型并不是机器学习的终极目标，其最终旨趣是辅助人类分析和决策。

从运作流程来看，机器学习首先需要一个“数据集”（data set），其中记录着关于对象或事件的表现、特征或属性等信息，从数据中获得模型的过程就是“学习”或“训练”（training）。在这一过程中，机器通过执行某种算法来生成相应的模型，目标任务不同，使用算法也就不同，学得模型也会有所差异。一般而言，模型的精准度需要“投喂”大量的“训练数据”（training data）来获得，从而不断逼近真相；在学得模型之后，还需要使用测试集（test set）来测试其性能，要求它有一种“泛化”（generalization）能力，并在此基础上不断修正、改进或增强。

按照学习方式来说，它又主要分为监督学习（Supervised learning）、无监督学习（Unsupervised learning）、半监督学习（Semi-supervised learning）、强化学习（Reinforcement learning）。首先，监督学习是指机器在学习过程中有任务监督，任务是指算法通过观察学习已知输入和输出之对应关系的实例（这些实例中的训练数据是被集中标记过的），从而让机器掌握这一对关系的函数规律。一旦算法得到训练，它就能够将所学知识应用于预测不同（目标）数据集的正确答案。（[9]，p.16）（[10]，p.4）分类（classification）和回归（regression）是常见的监督学习。分类是试图找出描述和区分数据类或概念的模型，以便预测那些（未知标签的）对象的标签，其导出模型是基于对训练数据集的分析。（[9]，pp.12–13）分类算法的输出标签是离散、无序的，具有有限的结果集。与之不同，回归分析是通过统计学方法预测缺失或难以获得的数据值，同时它能够基于可用数据识别数据的分布趋势，因此回归算法的输出具有连续性。

其次，无监督学习恰恰相反，它意味着学习过程没有任务监督，训练数据集没有被标

①下文内容如无特殊标注，所说的“算法”皆指“机器学习算法”这个子类。

记，学习过程无需人类干预，纯粹是数据驱动的过程（data-driven process），让机器自动发现隐藏或潜在的关联，识别有意义的趋势和结构。[11]常见的无监督学习有聚类分析、关联规则学习等。聚类分析是一个把数据对象划分成子集——簇（cluster）的过程。每个簇中的对象彼此相似，并与其他簇中的对象相异，由聚类分析产生的簇的集合称为一个聚类。聚类的作用在于它可能发现数据内事先未知的群组。（[9]，p.288）关联规则学习是基于规则逻辑方法，用于在变量之间的大型数据集中发现它们之间的关系，即“if-then”语句。[11]，[12]比如A购买了一辆汽车（一个项目），那么他很可能会购买相应的保险（另一个项目）。当然，还有一种介于监督学习与无监督学习之间的算法，即半监督学习，顾名思义，它是指在学习模型时同时使用标记的和未标记数据进行训练。

最后，强化学习则是基于与环境交互构建自主（物理或虚拟）机器智能体（Agent）的一般框架，这些系统可以在无监督的情况下做出决策以执行特定任务。[13]强化学习包括智能体、环境、奖励和政策这四个要素，其运行轨迹也被称为马尔可夫决策过程（Markov Decision Process）。智能体在没有监督、不断试错的情况下，自动评估特定情境或环境中的最优行为或策略，即环境驱动（environment-driven approach）；同时，强化学习的系统和训练过程会通过设置奖励函数获取反馈，即智能体需要学习如何根据其观察结果和奖励反馈来选择好的行为，通过反复实验和测试努力实现奖励的最大化。

当然，以上仅是一些粗略的划分，并不完整。需要单独指出的是近些年来热门的深度学习（Deep Learning）算法，它是机器学习的子集，基于神经网络（Neural Networks）来建模和解决更为复杂的任务。深度学习通过几个处理层（输入层、隐藏层和输出层）从数据中学习并改进自己的性能。与传统机器学习相比，它的优势在于，随着数据量的增加，其性能比前者更为优秀。[11]这是由于它的哲学基础是联结主义（connectionism）：尽管机器学习模型中的单个生物性的神经元或单个特征并不智能，但大量的神经元或特征共同作用就可以表现出智能的行为。（[14]，p.443）因此深度学习的核心是通过多层的非线性变换从数据中提取高层次的特征，但它本质上还是机器学习。

三、机器学习算法的功能特征和社会应用

通过以上概述，我们大致了解了它的运作逻辑和机制。从技术性质来看，算法主要有特征抓取（feature extraction）、预测分析（predictive analytics）和智能决策（intelligent decision-making）这几种功能，它们相辅相成，在运行时交织在一起，最终形成了一套综合的社会治理（social governance）模式，作用和影响与日俱增。

第一，机器学习能够对数据进行特征抓取。它通过从数据中提取有效信息，构建可泛化的模型，从而总结特征。首先是预处理，它在海量的原始数据中清洗异常数据，并通过降维（dimensionality reduction）消除与任务目标无关的冗余数据，减少计算复杂度；同时，机器通过执行回归分析等算法来插补缺失或遗漏的数据。其次，机器学习能够增强数据的可分性，让同类数据在特征空间内聚集，分离异质性数据，这一操作是根据任务目标而定。最后，算法根据聚集的数据类别抓取其特征。这一功能在电子商务领域应用较广，比如公司经常基于消费者的历史消费数据精准地了解后者偏好，从而为其制定生产和销售战略提供参考。当然，算法所处理的也不局限于数字信息，它还能识别图像、语言等其他模式的数据。

第二，机器学习也有预测分析功能。与特征抓取一样，预测分析也同样依赖于数据，基于第一步，算法通过构建数学模型，能够预测和推断未来事件的发展趋势和结果，也即“预测建模”（predictive modeling）。就流程而言，预测似乎很好实现：选择一种建模技术，输入数据，然后生成预测。但这种方法生成的结果却不一定是可靠可信的，而要达到这一点，我们必须先理解数据和建模的目标，只

有完成这一步之后，再进行数据预处理和分割，最终才能构建、评估和选择模型。（[15]，p.26）所以反向而言，预测模型的失败因素主要有：（1）数据预处理不足；（2）模型验证不足；（3）不合理套用（比如将模型应用于此模型从未见过的空间中的数据）；（4）将模型过拟合（overfitting）到现有数据当中①。（[15]，p.3）预测功能在保险行业应用较广，比如A为私人汽车购买保险，此时保险公司就会利用算法根据A的信息评估风险并计算费用。在分析过程中，算法主要基于与A相关的历史数据及其特征，如驾龄、性别、历史出险次数等要素来预测其未来风险系数，并直接影响保费的高低。

第三，机器学习还能够进行智能决策。无论我们使用算法是为了抓取数据特征，还是预测事件趋势，本质上还是为了辅助决策。正如上述保险案例，保险公司基于对客户的特征抓取以及风险预测，最终目的是为了帮助公司给出一个合理方案。因此，智能决策是算法通过数据建模将不确定性转化为可量化的行动策略，在复杂的环境中根据目标任务自动或半自动做出“最优”的选择，模拟或辅助人类决策；当然，决策的准确性是算法具有粘性的关键因素。[16]从算法决策的角色程度来看，它或是辅助我们决策，或是直接替代我们决策。[17]从决策逻辑的类型来看，主要有规则驱动、数据驱动和混合增强决策。规则驱动决策是基于预先定义的规则逻辑来做出决策，数据驱动决策完全依赖于数据建模，混合增强决策则是同时结合两者。而从决策时序来看，主要有静态决策、序列决策和实时流式决策。静态决策是单次输入产生固定的决策，序列决策是在动态的环境下连续决策，实时流式决策则是毫秒级的响应。

总之，机器学习不仅功能完备，而且社会应用性极强。同时，它的应用也绝非单点式、片段式，而是具有连续性和广泛性，这三种功能交织形成的是一套综合的、全景式的社会治理模式②。这种模式在医疗领域得到了普遍践行，最经典的莫过于现代“专家系统”（expert systems）。与传统模式仅仅是一种“知识的综合”不同，现代系统其实包含了一组演绎规则数据库，通过这些规则——给定一组已知事实——可以推断出某些后果。在医学诊断中，专家系统读取和处理病人的医疗信息，根据其临床症状精准定位到其病情特征（特征抓取），而某些症状的出现还会触发系统提出或推测与这些症状相符的疾病（预测分析），并给出相应的诊断和治疗建议（智能决策）。具言之，算法能够综合信息给出疾病的评级量表，自动推断其内部参数的设置以及哪些特征会影响对原始标签的最准确预测，[18]并提供一套治疗方案。鉴于当下医疗数据普遍地以数字方式收集，也致使其更易于通过算法进行分析，由此可见，机器学习算法的社会治理模式在医疗领域已经渐趋成熟，当然，在其他领域中发挥的作用也不遑多让，它有能力为人类提供全方位辅助。

四、无可逃避的“先天技术缺陷”

由上可知，算法在社会中无孔不入，但它在给人类带来便利的同时也引发了一些伦理风险和危机，如隐私泄露、算法歧视、归责模糊等。据此，有学者批判算法的这种非伦理性，提出要在伦理意义上限制或改造算法以规避这些问题，比如要求公开算法的源代码和设计架构、选取合理的数据集等措施。然而，当我们从技术原理切入，可以察觉到这种批评和改造方案本质上是一种不切实际的“苛求”。因为算法在运行时存在着一些无可逃避的“缺陷”，如数据依赖、算法偏见和算法黑箱，它们并非如我们所愿，可以通过对算法的伦理改造来克服，而是算法技术自身的先天不足，因此导致

①过拟合是指这样一种现象：模型对于用于训练的样本数据拟合得过好，反而导致在新的测试数据集上表现不佳。其对立面是欠拟合（underfitting），它指的是由于模型不合适（如过于简单），从而不能在训练数据上较好地拟合。具体而言，如果把样本数据作为模型训练的经验，那么模型的训练结果就和样本数据的选取关系很大。假如样本不够充足或不具有代表性，而模型又过于复杂，就很容易让模型学习到与任务无关的特征，并且和任务错误地建立联系。

②“治理”概念此处并非狭义地指政府的行为方式，而是广义地理解为在某一领域进行连续性、贯通性的操作模式。

的伦理风险也同样是无可逃避的。这一论断或许有违认知常识，但从算法的运作逻辑和功能特征来看，所言非虚。

第一，数据依赖（data dependency）是机器学习算法的“阿克琉斯之踵”。它既是机器学习的首要特征与核心优势，也是机器学习与传统算法最大的差异。传统算法的特点是模型是确定的，与数据无关，只要给一个输入，在算法中进行一些既定的操作，最终就能得到输出结果；但机器学习的模型不是直接给定、不可更改的，而是从数据中学习得来的，并且算法利用学得的模型可以对新的同类数据进行操作，得到对应的结果。[19]由此可见，机器学习相较于传统算法的固定性，可以凭借对数据的学习不断提高性能。但所谓“成也萧何、败也萧何”，数据同时也是机器学习的“缺陷”，后者在输入前提、模型训练、性能测试、结果输出方面，整个运作程序的好坏都取决于数据，可以说，算法无数据则空，数据无算法则盲。没有数据，算法仅仅是某种惰性的、无意义的机器，[20]更妄谈执行监督学习、无监督学习之类的操作了。

具体而言，机器学习对数据的依赖性主要有数量、质量和分布三种，它们的影响贯穿算法学习始终。首先，数据数量不仅是算法学习模型的前提，也是模型发挥功能的依赖。一方面，模型的获得就需要大量的训练数据，另一方面，机器想要应用特征抓取等功能，只有数据越多，特征把握才会越具体，预测分析才会更准确。尤其是深度学习，只有给它“投喂”越庞大的数据集，它的学习模型及其推理能力才会更强大，所以在“大数据时代”机器学习更加容易”。其次，数据质量也会影响到学习模型的性能和精度，计算机学界流传着一句格言：“垃圾入，垃圾出”（garbage in, garbage out），即如果进入训练的数据存在质量问题，那么程序基于这些数据做出的概括也会有问题。[21]比如机器在无监督学习时，如果给出的是一些虚假或劣质数据，那么它通过聚类分析只能发现一些错误的结构或关联，从而生成无意义的结果。最后是数据的分布依赖性，由于许多算法中的一个主要假设是训练数据和未来数据必须在相同的特征空间中并且具有相同的分布，[22]因此如果训练数据与实际应用场景中的数据分布不一致时，模型的性能会显著下降，其效果类似于“刻舟求剑”。例如，汽车企业在利用强化学习训练自动驾驶模型时，如果只在高速公路上训练，但实际应用场景却是在城区，那么设置的奖励函数其实是错位的，模型效果就远不如预期。由此可见，机器学习算法的性能在很大程度上取决于数据的数量、质量和分布，也就不可避免地产生了“数据依赖性”，而这一“缺陷”是现阶段无法在技术上克服的。

第二，算法偏见（algorithmic bias）也是一种先天缺陷。就概念内涵而言，偏见有两种性质，其一是积极的，主要是指一种偏好性的认知状态，比如偏爱某个对象；这类偏见大部分在认知上可靠且在道德（和规范）上中性。[21]其二是负面的，是指一个人对特定个人或群体的否定的信念或行为。[23]，[24]从偏见类型来看，又有显性偏见和隐性偏见；显性偏见是指人对自己的某种认知状态有明确的意识可及性，而隐性偏见则是一种潜意识或无意识的状态。需要注意的是，偏见类型与内涵是相互交错的，既有积极的显性偏见，也有积极的隐性偏见。将偏见移置到算法的语境中，考量其致因要素，也有三个来源：其一，数据的偏见，刚刚已经提到，算法具有极强的数据依赖性，如果机器使用的数据存在偏见，这种偏见的影响就会在学习过程中被接受并扩大，从而导致输出决策也具有偏见。其二，算法开发或设计者的偏见，无论他们是否有意将某种认知倾向植入算法设计当中，都会引发偏见性输出。一旦这种偏见是负面的，将不可避免地导致算法歧视等伦理问题。其三，算法自身的偏见，这也是“先天”的偏见；严格说来，对算法的使用就是一种偏见，退而言之，即使我们保证了数据的无偏性和中立性，但从算法对数据的处理和加工步骤来看，它的运作就是以偏见为前提。以聚类分析为例，算法在将数据划分成不同的“簇”时会经过优先级排序（prioritization）、分类、关联（association）和过滤（filtering）等一系列

操作，在优先级排序中，开发者会制定一套计算并用于通过排序程序定义排名的标准，这些标准就是一种偏好，它以牺牲其他事物为代价强调对某些事物的关注。[25] 换言之，不管排序标准是有意设定还是无意学得，它总归是一种倾向，而分类、关联和过滤操作都有此类问题。所以无论从人类还是算法的角度，偏见问题都是无法完全消除和克服的一个特征。[24]

因此，偏见不仅是算法的内在属性，更是其设计的逻辑前提。算法偏见本质上是在数学或统计学意义上形成的技术性产物。[26] 机器在学习过程中，为了明确和学习目标任务，它们必须比其他特征空间的数据元素更重视某些数据元素，换言之，它们必须对世界的秩序有所假设。尽管社会普遍呼吁算法应当去除偏见，但事实截然相反，它们在现实世界中运作时必须从其应用环境中提取假设，以适应和学习。于是根本而言，算法永远不能是中立的，它明确需要偏见和假设。[27] 这一点在计算机教科书中也明确指出："没有假设就没有推理和预测。"[28] 归根结底，大部分算法偏见在认知上可靠，在道德或伦理（规范）上中立，仅仅是技术需要的结果。

第三，算法黑箱是机器学习尤其是深度学习面临的又一个缺陷。算法黑箱主要是指机器学习过程的不透明性（opacity），不透明性是指即使我们知道算法输入的变量和输出的预测结果，但我们仍然无法知道模型内部是如何运作的，也不知道它的决策逻辑。[29] 与之相反的是算法的透明性（transparency），主要表现为两个层次，一是可理解性（Interpretability），二是可说明性（Explainability）①。可理解性是指观察者能够理解决策原因的程度，而可说明性则是指观察者可以获得说明的一种模式。[30] 换言之，可理解性偏向于模型自身的结构或参数是否天然具备可被人类理解的特性；而可说明性则是指通过外部方法或工具对模型的预测结果生成事后说明。二者区别在于，可说明性并不一定要完全忠于初始模型计算或决策过程的说明。[31] 所以算法黑箱并不仅仅意味着算法模型的决策复杂性，即人类通常无法追溯其决策依据，还意味着这样一种状态：即使开发（设计）者公开所有的源代码或架构等信息，也不一定能够提供一套合理的说明步骤。因为仅仅有可理解性是不够的，许多非专业人士并没有足够的知识储备，因此还需要对黑箱的说明，即用简单的、有意义的词汇来描述算法做出如此决策的原因。[32] 由此可见，可说明性比可理解性的要求更高，二者的概念内涵是互补的，共同构成了算法的"可解释性"。[33]

当然，算法黑箱的成因也是多元的，其中在伦理上中立的因素主要有：其一，商业机密与安全限制；很多企业为了保护知识产权与防止恶意攻击，会隐藏涉密的模型细节。其二，语言的特殊性；编写（和阅读）代码以及算法设计是一项专业技能，代码需要用特定的编程语言编写，这与人类语言相当不同，对大多数人来说不仅较少接触，还难以理解。其三，训练过程不透明；机器学习在训练时可能会根据庞大的数据集自动调整参数，以环境驱动为特征的强化学习就是如此，但人类却很难察觉到这一点。其四，模型复杂度过高，现代机器学习尤其是深度学习模型通过多层非线性结构捕捉数据特征，参数规模庞大，导致决策逻辑难以直观追溯。其五，算法开发规模庞大；这主要是由于算法（如谷歌搜索引擎的底层算法）通常是由团队构建的多组件系统，即使是内部人员也未必清楚地了解全貌。[34] 从这些原因可以看出，在多数情况下，算法黑箱并不是开发（设计）者有意为之，而是技术的局限性，以我们目前的技术水平既不能完全理解黑箱，也无法给之提供一个常识性说明，因此也是一种"先天缺陷"。有鉴于此，当下很多学者也开始转变思路，不再去尝试解释黑箱模型，而是去创建一个可解释的算法模型。[35]

总之，与我们习以为常的恰恰相反，数据

①一直以来，学界对于"interpretability"和"explainability"这两个词汇是混用的，都翻译为"可解释性"，但实际上二者之间存在一定的差异，笔者认为，二者的概念内涵是互补的，它们"合体"才可以译为"可解释性"。

依赖、算法偏见和算法黑箱是机器学习在开发设计和应用过程中无可逃避的技术缺陷，而非使用伦理手段可以克服。从伦理视角出发而提出对算法自身的技术治理设想是几乎没有意义的，只能针对其导致的伦理后果做出补救或弱化措施。

余　论

经过对机器学习算法术语的澄清、技术原理的拆解以及运作机制的梳理，我们应当重新思考算法伦理的研究模式。以往我们过于注重伦理的规范作用，从而提出一些不切实际的措施来限制或约束技术来拟合我们对合伦理性的算法形态的构想，殊不知这些缺陷并非是由我们缺乏伦理关照而导致的；相反，它们是“前伦理”的形态，是技术自身局限的产物，由其应用导致的社会问题才具备伦理属性。因此，对算法伦理中数据依赖、偏见和黑箱等问题的研究重心应当着眼于如何用伦理手段治理它导致的反伦理后果，而非一厢情愿地用伦理介入将技术改造为理想的形态。换言之，我们需要接受算法的不完美性，在进行伦理批判时保留一种同情心理，尽量不去苛责算法自身的局限性。

当然，我们也不必过于悲观：既然算法技术暂时无法提高，难道我们要深陷于这些由它们导致的伦理困境之中无法更改现状吗？答案是否定的。一方面我们固然需要时间来不断改进技术完善它的可操作性，力求未来在根本上规避这些缺陷；但另一方面，我们不妨换一种思路，以“伦理”治理“伦理”，即用伦理措施来弱化或补救算法的先天技术缺陷导致的伦理后果。这一进路正逐渐得到学界的关注和讨论，比如荷兰学者斯考特·罗宾斯（Scott Robbins）提出“封装”（envelopment）的伦理设计，其核心思路是在某种程度上限制系统，使其能够在有限的容量下实现所需的输出。[36] 简言之，它是在算法周边加筑“防线”，即通过约束算法的运行环境和运行结果（如使用场景或物理空间、输入输出范围）来避免不可控的风险，保证其输出合于伦理，而无需考虑在这一过程中算法是否有偏见、黑箱是否可解释等先天缺陷①。总之，算法伦理研究固然要对其伦理风险有所警惕，但也要尊重和基于技术自身的局限性，公允地进行批判和治理。

［参 考 文 献］

[1] 孙保学. 人工智能算法伦理及其风险 [J]. 哲学动态, 2019,(10): 93–99.

[2] 高斯扬. 算法伦理的类型分析——基于“技术–伦理”框架 [J]. 科学学研究, 2024, 42（9）: 1808–1815.

[3] 夏明月、陈冬阳. 机器学习伦理风险及其防控治理 [J]. 山东社会科学, 2024,(8): 106–114.

[4] 李伦、张晓燕. 算法不公正：问题类型与治理路径 [J]. 自然辩证法通讯, 2025, 47（5）: 76–83.

[5] 闫瑞峰. 算法设计伦理治理的立场、争论与对策 [J]. 自然辩证法通讯, 2023, 45（6）: 1–9.

[6] 孟令宇. 从算法偏见到算法歧视：算法歧视的责任问题探究 [J]. 东北大学学报(社会科学版), 2022, 24(1): 1–9.

[7] Mitchell, T. M. *Machine Learning*[M]. New York: McGraw-Hill Science/Engineering/Math, 1997, 2.

[8] 周志华. 机器学习 [M]. 北京：清华大学出版社, 2016.

[9] Han, J. W., Kamber, M., Pei, J. 等. 数据挖掘概念与技术 [M]. 范明、孟小峰 译, 北京：机械工业出版社, 2012.

[10] Liao, S. M. *Ethics of Artificial Intelligence*[M]. New York: Oxford University Press, 2020, 4.

[11] Sarker, I. H. 'Machine Learning: Algorithms, Real-world Applications and Research Directions'[J]. *SN Computer Science*, 2021, 2(3): 160.

[12] Agrawal, R., Imieliński, T., Swami, A. 'Mining Association Rules Between Sets of Items in Large Databases'[A], Buneman, P., Jajodia, S. (Eds.) *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*[C], New York: Association for Computing Machinery, 1993, 207–216.

[13] Buffet, O., Pietquin, O., Weng, P. 'Reinforcement Learning'[A], Marquis, P., Papini, O., Prade, H. (Eds.) *A Guided Tour of Artificial Intelligence Research: Volume I: Knowledge Representation, Reasoning and Learning*[C], Cham: Springer, 2020, 389–414.

[14] Goodfellow, I., Bengio, Y., Courville, A., et al. *Deep Learning*[M]. Cambridge: The MIT Press, 2016.

①私以为用伦理“封装”算法是目前最具可行性的措施之一，不过限于篇幅在此只能做一个简单的展望。

[15]Kuhn, M., Johnson, K. *Applied Predictive Modeling*[M]. New York: Springer, 2013.

[16]Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., et al. 'What Influences Algorithmic Decision-making? A Systematic Literature Review on Algorithm Aversion'[J]. *Technological Forecasting and Social Change*, 2022, 175: 121390.

[17]Edwards, J. S., Duan, Y., Robins, P. C. 'An Analysis of Expert Systems for Business Decision Making at Different Levels and in Different Roles'[J]. *European Journal of Information Systems*, 2000, 9(1): 36–46.

[18]Grote, T., Berens, P. 'On the Ethics of Algorithmic Decision-Making in Healthcare'[J]. *Journal of Medical Ethics*, 2019, 46(3): 205–211.

[19]贾壮. 机器学习与深度学习[M]. 北京：北京大学出版社，2020, 7.

[20]Gillespie, T. 'The Relevance of Algorithms'[A], Gillespie, T., Boczkowski, P., Foot, K. (Eds.) *Media Technologies: Essays on Communication, Materiality, and Society*[C], Cambridge: The MIT Press, 2014, 167–194.

[21]Johnson, G. M. 'Algorithmic Bias: On the Implicit Biases of Social Technology'[J]. *Synthese*, 2021, 198(10): 9941–9961.

[22]Pan, S. J., Yang, Q. 'A Survey on Transfer Learning'[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(10): 1345–1359.

[23]Howard, A., Borenstein, J. 'The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity'[J]. *Science and Engineering Ethics*, 2018, 24(5): 1521–1536.

[24]Howard, A., Borenstein, J. 'Hacking the Human Bias in Robotics'[J]. *ACM Transactions on Human-Robot Interaction (THRI)*, 2018, 7(1): 1–3.

[25]Diakopoulos, N. 'Algorithmic Accountability: Journalistic Investigation of Computational Power Structures'[J]. *Digital Journalism*, 2015, 3(3): 398–415.

[26]Vicente, L., Matute, H. 'Humans Inherit Artificial Intelligence Biases'[J]. *Scientific Reports*, 2023, 13(1): 15737.

[27]Amoore, L. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*[M]. North Carolina: Duke University Press, 2020, 74–75.

[28]MacKay, D. J. C. *Information Theory, Inference and Learning Algorithms*[M]. Cambridge: Cambridge University Press, 2003, 345.

[29]Staartjes, V. E., Kernbach, J. M. 'Foundations of Machine Learning-Based Clinical Prediction Modeling: Part III—Model Evaluation and Other Points of Significance'[A], Staartjes, V. E., Regli, L., Serra, C. (Eds.) *Machine Learning in Clinical Neuroscience*[C], Cham: Springer, 2022, 23–31.

[30]Miller, T. 'Explanation in Artificial Intelligence: Insights from the Social Sciences'[J]. *Artificial Intelligence*, 2019, 267: 1–38.

[31]Tsang, W. K., Benoit, D. F. 'Interpretability and Explainability in Machine Learning'[A], Ohsawa, Y. (Ed.) *Living Beyond Data*[C], Cham: Springer, 2023, 89–100.

[32]Gilpin, L. H., Bau, D., Yuan, B. Z., et al. 'Explaining Explanations: An Overview of Interpretability of Machine Learning'[A], *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*[C], Turin: IEEE, 2018, 80–89.

[33]Leblanc, B., Germain, P. 'On the Relationship Between Interpretability and Explainability in Machine Learning'[J]. arXiv preprint, arXiv: 2311.11491, 2023.

[34]Burrell, J. 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms'[J]. *Big Data & Society*, 2016, 3(1): 1–12.

[35]Rudin, C. 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead'[J]. *Nature Machine Intelligence*, 2019, 1(5): 206–215.

[36]Robbins, S. 'AI and the Path to Envelopment: Knowledge as a First Step Towards the Responsible Regulation and Use of AI-powered Machines'[J]. *AI & Society*, 2019, 35(2): 391–400.

[责任编辑　李斌]