技术性正当程序：人工智能时代程序法和算法的双重变奏

刘东亮*

摘　要：人工智能的广泛运用正在重塑政府的运作方式，现代社会的权力形态和权力结构均发生改变，算法权力悄然兴起。与此同时，传统行政程序在人工智能的适用场景中失去了原有的效用，诸如“听取意见”、“说明理由”等权利保障措施，对于瞬间即作出决定的自动化机器几无适用的余地，这意味着行政法的关注焦点需要从传统行政程序向计算机程序算法拓展。对算法权力的规制和监督，需要从算法设计的源头构建“技术性正当程序”，即通过程序的代码化实现下列要求：算法公开、透明并具有程序一致性；算法具有可解释性，能提供决策的相关逻辑和实质性信息；对决策结果允许质疑，在专业人员协助下审查算法，有错误及时修正，等等。技术性正当程序要求行政法与时俱进，跟得上技术发展的步伐。

关键词：人工智能；算法权力；算法设计；技术性正当程序

近年来，每个人都能切身感受到，人工智能已经走进我们的日常生活。人工智能技术的广泛应用，在给人们带来新奇和兴奋感的同时，也带来了深深的焦虑与不安。因为，没有人确切知道，人工智能的快速发展，会导致哪些未知的风险——风险与焦虑总是如影随形、相伴而生。〔1〕故而，我们必须直面一系列由人工智能引发的既已存在的现实挑战。

一、从行政程序到程序算法：人工智能时代行政法关注对象的变化与拓展

过去，在人工智能广泛渗入行政领域之前，行政法关注的核心问题，除了行政诉讼即行政程序。这一特征在英美法系尤为明显。美国已故著名行政法学家施瓦茨教授曾言：“相较于实体法，行政法更多的是关于程序和救济的法律。”〔2〕大陆法系对行政程序的重视程度，尽管不能与英美法系同日而语，但从1960年代以来，德、日等国对行政程序的重视亦与日俱增。〔3〕我国虽然迄今尚未出台统一的行政程序法典，但1989年颁布的《行政诉讼法》早已将“违反法定程序”作为撤销（具体）行政行为的理由之一；1996年颁布的《行政处罚法》分别规定了作出处罚决定的简易程序、一般程序和听证程序。在新中国的立法史上，这些规定具有划时代的意义。在学术方面，自1993年季卫东教授发表《法律程序的意义》以

*　西安交通大学法学院教授，法学博士。本文系国家社科基金后期资助项目“正当程序的法理：法律和社会科学多视角的分析”（19FFXB001）的阶段性成果。本文所涉技术性问题，得到西安交通大学人居环境与建筑工程学院赵腾远研究员、数学与统计学院孟德宇教授和李丽敏教授等多位专家的帮助，特此致谢！

〔1〕　*See* Daniel J. Solove & Danielle K. Citron, *Risk and Anxiety: A Theory of Data-Breach Harms*, 96 Texas Law Review 737, 786 (2018).

〔2〕　Bernard Schwartz, *Administrative Law* 3 (Little, Brown & Company 1976).

〔3〕　参见姜明安：《21世纪中外行政程序法发展述评》，载《比较法研究》2019年第6期，第130–143页。

来,[4]学界对程序法的研究高潮迭起。1997年,曾有行政法学者提出了“法即程序”的命题;[5]其后,又有很多学者指出,法治就是“程序之治”。[6]在实践方面,从1998年法院在“田永诉北京科技大学案”中首次运用正当程序的精神判案以来,对正当程序的司法适用不断向纵深发展。可以毫不夸张地说,对行政程序特别是正当程序的研究已站到了行政法学的风口浪尖。

然而,不难发现,原本高举高打的行政程序在人工智能的适用场景中即刻失去了原有的效用。例如,正当程序所强调的“听取意见”、“说明理由”等权利保障措施,对于瞬间即作出决定的自动化机器几无适用的余地。正因为如此,有些国家的行政程序法干脆“顺水推舟”,明确规定采取自动化行政时免除事先听证和说明理由,从而导致当事人的程序性权利被完全架空。[7]简言之,传统行政程序在人工智能的适用场景中遭遇了“滑铁卢”。

面对传统行政程序的挫败,我们不得不将目光投向行政领域中运用得越来越多的自动化决策系统(automated decision systems),[8]审视这些形形色色的智能机器当中究竟发生了什么。目前,与司法领域能否由“阿尔法法官”(Judge Alpha)作出裁判尚有争议的情况相比,在行政领域,由智能机器接棒人类作出大量行政决定已是显而易见的事实。当下,我国的“互联网+政务”正如火如荼地全面推进。尤其是,我国在税务行政中已率先实现了全自动行政行为。在现实生活中,我们也都听说过或者亲身经历过购车(房)摇号、摇号入学、电子罚单、“落户秒批”、高校利用大数据分析识别贫困生并给其发放补贴,等等。特别是2020年春,我国在疫情防控中广泛使用的“健康码”、“一码通”,都是行政自动化决策的鲜明例证。

我们知道,这些形形色色、面貌各异的行政自动化决策系统其载体均为计算机。计算机的工作过程就是执行程序——只是此“程序”(program)非彼“程序”(procedure)。计算机程序是由程序开发人员使用某种程序设计语言编写的以代码形式表示的能够为计算机识别并予以执行的指令集合。计算机程序和法律程序存在明显差异,但二者也有共同点,它们都是实现预定目标的一系列步骤,可以通过代码予以表达,即“程序的代码化”。网络法学界有一句人人皆知的名言“代码即法律”(Code is law)。[9]实际上,这句话反过来也能成立。法律,特别是程序法规则,能够很方便地编译到代码中去。

根据计算机科学界的理解,计算机程序的核心是“算法”。著名瑞士计算机科学家沃思(Niklaus Wirth)曾经提出一个计算机领域人人皆知的名言:算法+数据结构=程序(Algorithm+Data Structures = Programs)。该公式是沃思1976年出版的一部专著的书名。[10]很多人认为,该公式对计算机科学的影响足以媲美物理学中爱因斯坦的质能公式“$E=mc^2$”,因为该公式简洁而又充分地说明了程序的本质,沃思亦因此获得1984年图灵奖。在计算机科学界,还有一个广为人知的说法:数据是“燃料”,算法是“引擎”。易言之,算法是程序的灵魂。有人甚至认为,计算机科学就是研究关于“算法”的学问。[11]

那么,究竟什么是算法呢?关于算法的定义五花八门,其词源流变也相当复杂。据称,“算法”(algorithm)一词源于9世纪一位波斯数学家“Al-Khowârizmî”的名字。[12]不过,算法的产生远早于这位数学家;算法不限于数学,也不限于计算机。从石器时代开始,算法就已经是人类生活的一部分了。[13]

〔4〕 参见季卫东:《法律程序的意义》,载《中国社会科学》1993年第1期,第83-103页。

〔5〕 肖凤城:《论“法即程序”——兼论行政程序法的重要性》,载《行政法学研究》1997年第1期,第4-7页。肖凤城教授其后又在《行政法学研究》发表《再论“法即程序”》《三论“法即程序”》,出处从略。

〔6〕 参见梁捷:《法治必须强调程序正义》,载《光明日报》2013年7月10日,第2版。

〔7〕 参见[德]毛雷尔:《行政法学总论》,高家伟译,法律出版社2000年版,第443页。

〔8〕 本文所称“自动化决策”实际上是指“自动化决定”。“决策”(decision)一词系沿用管理科学决策理论学派的概念(“管理就是决策”),以便于我们从多学科的视角审视智能机器的“决定”。

〔9〕 参见[美]莱斯格:《代码2.0:网络空间中的法律》,李旭、沈伟伟译,清华大学出版社2018年版,第6页。

〔10〕 Niklaus Wirth, *Algorithms + Data Structures = Programs* (Prentice-Hall 1976).

〔11〕 参见周志华:《机器学习》,清华大学出版社2016年版,第1页。

〔12〕 Donald E. Knuth, *The Art of Computer Programming: Fundamental Algorithms (Volume 1)* 2 (Addison-Wesley 1997).

〔13〕 参见[美]克里斯汀、格里菲思:《算法之美》,万慧、胡小锐译,中信出版社2018年版,引言。

可以说,算法就是解决问题的方法,是对解题方案准确、完整的描述。而将解决实际问题的方法变成计算机可以运行的程序,中间的桥梁就是计算机的算法。〔14〕

众所周知,自1956年达特茅斯会议麦卡锡(John McCarthy)首次提出“人工智能”(AI)的概念,六十多年来,人工智能的发展跌宕起伏,已经历了“两落三起”。最近这一波人工智能浪潮的兴起,主要得益于大数据的积累及硬件技术进步带来的计算力之提升。其中,虽然大数据起着推波助澜的作用,但作出那些至关重要的决策并不取决于数据本身,而在于如何对数据进行算法上的分析。〔15〕在人工智能时代,由源源不断的数据驱动的算法已成为新的权力代理人。〔16〕无处不在的人工智能算法存在于人们的手机和个人电脑里,存在于政府机关、企业和公益机构的服务器上,存在于共有或者私有的云端之中。虽然我们不一定能时时刻刻感知到算法的存在,但人工智能算法已高度渗透到我们的生活之中。并且,随着人工智能技术的不断成熟,这种渗透还会进一步加速,从而日益深刻地改变人类生活的方方面面。〔17〕因而,如何对算法及通过算法获得或攫取的权力进行有效监督,防止算法通过各种看不见的“算计”暗中侵害我们的权利,是人工智能时代行政法学面临的严峻问题。

有必要指出,行政法需要关注的算法,不仅仅是指行政机关实施自动化决策时所运用的算法,还包括那些对社会有着巨大影响的私营机构使用的算法。随着数字经济的快速发展,很多网络商业平台的角色已经从单纯提供网络交易场所、撮合交易而升级进化为网络空间的秩序塑造者。社交平台也是如此。它们既制定平台参与规则,也负责执行规则,同时对发生的纠纷进行仲裁。这就意味着,这些传统意义上的市场监管对象,通过网络空间的架构方法(算法)攫取了相当大一部分公权力。在网络空间,超级平台权倾天下。它们有权决定谁有资格加盟自己的平台,哪个应用程序能在线上商店得到热捧,又有哪些应用程序有幸成为设备的默认出厂设置。超级平台可以成就一款应用程序,也能让它黯然离场。一言以蔽之,在互联网世界,站在食物链顶端的超级平台已成为游戏规则的制定者。〔18〕以谷歌公司为例,欧盟反垄断当局即认为,它并非市场中的一个竞争者,而是为他人设置竞争条款的权力核心和“造王者”(king-maker)。〔19〕

商业形态和市场结构的剧变导致“国家”与“市场”的界限日益模糊。不仅如此,政府与私营机构的关系也正在发生前所未有的变化。过去,政府与私营机构之间主要是监管关系。如今,政府和企业之间不仅存在交易关系,它们还会因利益交换的需要而进行隐秘的联姻。当然,“猫鼠结盟”是为了双方获利。〔20〕有美国学者对此评论说,政府和私营机构共同具有的黑箱结构和不断加强的合作关系使两者越来越为相像,并形象地称这些私营机构为“老大的马仔”(the big brother’s little helpers)。〔21〕

政府和私营机构的合作、联手进一步改变了现代社会的权力形态。在数字化时代,技术赋能促成了数字权力的诞生。早在20世纪90年代,麻省理工学院教授尼葛洛庞帝(Nicholas Negroponte)就预言了数字化社会的四大特征:去中心化、全球化、追求和谐、赋予权力。〔22〕技术赋权生成的权力形态正如同福柯(Michel Foucault)所描述的现代权力,它并非中央集权式的环状结构,而是错综复杂、多中心存在

〔14〕 参见周玉萍主编:《信息技术基础》,清华大学出版社2017年版,第108页;吴军:《数学之美》,人民邮电出版社2020年版,第323页。

〔15〕 Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* 21 (Harvard University Press 2015).

〔16〕 张凌寒:《风险防范下算法的监管路径研究》,载《交大法学》2018年第4期,第49页。

〔17〕 参见腾讯研究院等:《人工智能》,中国人民大学出版社2017年版,第63页。

〔18〕 参见[英]扎拉奇、[美]斯图克:《算法的陷阱》,余潇译,中信出版社2018年版,第229页。

〔19〕 Frank Pasquale, *supra* note 15, at 69.

〔20〕 以联邦快递(FedEx)为例,在其与美国政府开展合作之后,即获得了政府赋予的一系列特权和豁免权。华为公司的邮件多次被联邦快递无理扣押和“错递”,其背后的原因不言而喻。参见卢戈:《美媒曝:联邦快递错运事件与配合美政府打压华为有关》,载《环球时报》2019年6月5日,第3版。

〔21〕 Frank Pasquale, *supra* note 15, at 51–52.

〔22〕 Nicholas Negroponte, *Being Digital* 228 (Alfred A. Knopf, Inc. 1995).

的网状结构。[23]无处不在的数字权力就弥散于这张大网中。这种权力结构的“多中心存在”,表面上与“去中心化”相牴牾,实则互为成辅。因为,“去中心化”的网络社会,实际上还有一种“趋中心化”的趋势共存。以社交平台为例,平台已有的存量用户越多,就越能吸引更多新用户的加入,并借助用户的路径依赖心理产生“锁定效应”,进而获得在“超链接社会”(hyperlinked society)和“超联结社会”(hyper-connected society)才可能产生的“超级权力”,即利用用户集中、通过算法架构发挥更大的影响力,重新定义个人与个人、消费者与企业、公民与国家之间的关系。由于这种超级权力是通过数字化技术获得的权力,而数字化技术总要借助于某种算法才能够实现,因此,这种权力实质上就是算法权力(algorithmic power),或者至少可以说,其核心是算法权力。[24]

算法权力的兴起及其可能的异化风险,要求加强对算法的规制和监督。而要监督和规制算法,就必须深入了解算法,这是在当下的数字化、网络化和智能化社会,行政法学面临的“复杂性”挑战。为了有效应对这种挑战,行政法的关注焦点需要从传统行政程序向计算机程序算法进行拓展。

二、自动化决策系统和算法的工作原理及其局限性

有人工智能从业者声称:“算法唯一的工作是让你能够更加有效、更加精确地达成目标,而不是告诉你为什么。”[25]然而,我们认为,追问“为什么”非常重要。因为人工智能正在深刻影响人类生活,如果不对这种技术及其效用、风险等有所了解,就很难应对自动化决策的错误、不合理、不公正等问题,更不能给出一个适度实现技术有效性的规范框架。[26]

(一)自动化决策系统和算法的工作原理

1. 利用计算机解决问题的基本过程:从分析问题到算法设计

事实上,看似功能强大的计算机并不具备直接解决问题的能力,而必须通过执行人类为其编写的程序才能达到解题的目的。程序中的各种指令代码是人机交互的工具,能够把人类的解题方法传递给计算机,并由计算机予以执行。

利用计算机解决问题的基本过程,可概括为五个步骤:分析问题、设计算法、编写程序、调试运行、检测结果。其中,设计算法起着关键作用。设计算法需要在细致分析待解问题的基础上,抽象出相应的数学模型,确定使用的数据结构,然后设计实施各种操作指令。易言之,设计算法就是寻找解决问题的方案。只有找到适当的解题方案,得到某种解题的智慧,才能构建一台可表现智能行为的机器来完成某项工作。机器表现的智能等级受到算法转化的智慧所限制,算法设计的优劣也直接影响程序的通用性和有效性,影响到解决问题的效率。[27]

2. 算法的特征与描述方法

算法具有以下几项基本特征:(1)确定性。算法的每一个步骤都必须明确定义,不允许有歧义性和多义性。(2)有穷性。算法在执行有限步骤之后终止。算法必须在有限的时间内完成任务。(3)可行性。针对实际问题设计的算法,应当确保能得到满意的结果。正是这几项特征及其要求,决定了计算机不是万能的。对于那些不可计算、不可判定的问题,不可能设计出解决问题的算法。也就是说,计算机对有些问题无法处理。当然,计算机无解的问题,人类可通过其他手段予以解决。

在设计算法及事后解释算法时,需要对算法进行描述。常见的描述方法有:自然语言、流程图、伪代

〔23〕 福柯认为,现代权力是毛细血管状的,它不是从某个核心源泉发散出来的,而是遍布社会机体的每一个角落和最细小的末端。*See* Michel Foucault, *Power/Knowledge: Selected Interviews and Other Writings* 39 (Pantheon 1980).

〔24〕 关于“算法”的权力属性,参见Taina Bucher, *If…Then: Algorithmic Power and Politics* 3-8 (Oxford University Press 2018)。

〔25〕 转引自李彦宏等:《智能革命:迎接人工智能时代的社会、经济与文化变革》,中信出版社2017年版,第228页。

〔26〕 参见沈岿:《面对传统、现在与未来的行政法学》,载《行政法论丛》第24卷,法律出版社2019年版,“卷首语”。

〔27〕 参见周玉萍,《信息技术基础》,第107页以下。

码等。这几种常用的描述方法各有优缺点,在解释算法时可以结合使用,相互补充。〔28〕

3. 自动化决策何以可能:从“推理”到“预测”

在现实生活中,当人们遇到一个复杂的“非线性”问题时,总是千方百计地将其分解或(近似)转化为一系列简单的线性问题,然后逐个解决。〔29〕从算法设计的角度说,这种决策过程就是“化繁为简”,即将复杂的计算过程逐步归结为一系列简单过程的重复。当然,所谓“简单的重复”并不是单纯的罗列或再现,而是带有层次性,在程序上表现为循环,它仅仅是计算量的增加,而这正好是计算机的特长。

以使用较多的决策树算法为例,算法可以看作是由一个又一个的二元决策组成的巨大的决策树。事实上,我们做的所有事情,几乎都可以分解成一连串基于二进制输入的二元决策。因为“任何事物都是一分为二的”,世间万事万物都可作阴阳二分,并且这种二分过程可以延续到任意多层。〔30〕德国数学家莱布尼茨早在三百年前就已提出:生命也可以分解为一长串连续的二元决策。二元决策树可以为复杂的对象衍生出数百万甚至数十亿的节点。算法的决策树接收输入,然后将输入值带入方程和公式中运行,再将得到的答案作为新的输入,层层迭代演绎,创建出细节复杂的计算机指令字符。〔31〕

具体而言,在实践中运用相当广泛的专家系统(expert system),都是使用产生式规则(“if-then”)作出决策。可建立专家系统的领域一般都拥有大量的人类专门知识。程序设计者首先开发出存储了专家经验的知识库(知识库由许多专家的经验积累而成,它最终可达到一个专业顾问的知识水平并能超越任何单个人类专家),然后建立推理引擎和解释器等子系统。在运行专家系统时,一旦输入信息与产生式规则的前提条件相匹配,就会触发推理引擎执行推理,进而输出相应的决策结果。这种以知识为基础的专家系统可谓是人工智能领域最古老、最知名和最受欢迎的一种应用。在大数据时代到来之前,专家系统被视为人工智能技术最成功的例子。迄今为止,全世界已建立了数千个专家系统,用以解决工程、科学、医疗、军事、商业等各领域需要人类专家处理的复杂问题。〔32〕

传统的以知识为基础的专家系统在很多领域都有应用,它们给人类提供了极大帮助。但从严格意义上说,这些专家系统还不具有堪与人类匹敌的真正的“智能”,因为,它们都是由人类经“填鸭式教学”(手工输入知识)而成为“专家”的,其自身并不具备自主的学习能力。说得客气些,这些专家系统都只是一些高级的“书橱”。〔33〕

可以说,学习并不是专家系统的主要特征。但是,学习必须是任何可行的人工智能系统的一个组成部分。人工智能系统需要具备自己获取知识的能力,这种能力称为机器学习(machine learning)。早在1950年,有计算机科学之父和人工智能之父之称的阿兰·图灵(Alan Turing)在其划时代的论文《计算机器与智能》(Computing Machinery and Intelligence)中提出“图灵测试”的同时就提出了“机器学习”的可能性,并提出了“儿童程序”的设想。〔34〕因为,惟有具备学习能力才能克服人工智能发展的“知识瓶颈”。也只有使用学习方法而不是通过手工编码,才能让机器持续不断地有足够的数据可用,机器才可能拥有人类层次的“智能”。

20世纪80年代,被视为解决知识瓶颈之关键的机器学习,走上了人工智能的主舞台。1980年夏,在卡耐基梅隆大学举行了第一届机器学习研讨会。此后,机器学习蓬勃发展,各种机器学习技术百花竞放。尤其是,基于逻辑表示的符号主义(symbolism)学习和基于神经网络的联结主义(connectionism)

〔28〕 参见史巧硕、柴欣主编:《大学计算机基础与计算思维》,人民邮电出版社2015年版,第168-176页。

〔29〕 周玉萍,《信息技术基础》,第244页。

〔30〕 参见吴军,《数学之美》,第83页;王能超:《算法演化论》,高等教育出版社2008年版,第231页。

〔31〕 参见[美]斯坦纳:《算法帝国》,李筱莹译,人民邮电出版社2014年版,第14-15页。

〔32〕 参见[美]罗素、诺维格:《人工智能:一种现代的方法》,殷建平等译,清华大学出版社2013年版,第16页以下;[美]卢奇、科佩克:《人工智能》(第2版),林赐译,北京邮电出版社2018年版,第242页以下。

〔33〕 关于专家系统的优势和弱点,参见卢奇、科佩克,《人工智能》,第247-251页。

〔34〕 参见罗素、诺维格,《人工智能:一种现代的方法》,第17页。

学习，你追我赶，此消彼长。20世纪90年代，以支持向量机（SVM）为代表的“统计学习”（statistical learning）闪亮登场。到21世纪初，联结主义卷土重来，并掀起了以“深度学习”（deep learning）为名的热潮。人工智能技术也借助深度学习实现了一个新飞跃。〔35〕

与传统的依靠知识表示和推理作出决策的专家系统不同，包括深度学习在内的能够自行获取知识的各种机器学习算法，都是基于过去的数据并利用某种归纳偏好（inductive bias）来预测未来的趋势。任何一个有效的机器学习算法必有其归纳偏好，否则，它将被假设空间中看似在训练集上“等效”的假设所迷惑，而无法产生确定的学习结果。通过训练样本数据可使计算机产生某种归纳偏好，并产出它认为“正确”的（决策）模型。〔36〕

基于其建立的决策模型，运用机器学习方法的人工智能系统在接收新的信息和输入后，会拟合出一个映射函数，这个新的函数值既是对缺失信息的填补，也是对未来进行“预测”。可以说，使用机器学习和大数据技术的自动化决策系统，其决策结果在很大程度上是一种“预测”（prediction）。在大多数时候，我们很难相信这是预测，但从本质上说，它就是预测。〔37〕著名人工智能专家霍金斯（Jeff Hawkins）指出，智能的核心是对未来作出预测的能力。〔38〕诸葛亮之所以被中国民间奉为“智慧的化身”，盖因其“未出茅庐，已知三分天下”。我们或许并未意识到，预测是无所不在的。我们的生活中充满了预测。尤其是在复杂环境和不确定条件下，每个人都必须通过类比过去而预测未来。预测是一切决定的基础和关键。预测影响决策，进而影响行为。〔39〕不过，与人类决策不完全相同，人工智能系统通过分析过去的数据、填补缺失信息而预测未来。〔40〕这就意味着，除了借助规则推理方法进行问题求解，人工智能亦能通过分析已获得的数据、生成我们尚未掌握的信息（预测，或者叫函数拟合），尽可能分解、降低未来的不确定性而帮助人类作出决策。〔41〕正是在这个意义上，“机器能智能地行动”（弱人工智能假设）被大多数AI研究者认为是当然的。

4. 关于算法的两个特殊问题

在我国当前的“法律+人工智能”研究中，存在很多认识上的误区（需另外专文探讨）。限于篇幅，此处仅对与算法有关的两个特殊问题予以明确和澄清。

（1）算法是否“神秘”？

由于代码编写和算法设计在目前仍然属于专家技能，因此，很多非专业人士认为，算法非常“神秘”。事实上，算法虽然神奇（算法是人类智慧的结晶和体现），人们可以利用算法解决很多难题，但算法本身并不神秘。如前所述，算法就是解决问题的方法，是对解决问题的策略机制的描述。只有那些不公开、不透明被人为掩盖起来而显得“云遮雾罩”的算法才神秘兮兮。有些算法非常朴实简单，毫无神秘性可言。比如算盘被誉为最古老的数字计算机，珠算口诀可以说是最早的体系化的算法。〔42〕时代发展到今天，虽然人们在自动化决策系统中开发出了更为复杂的算法，但技术人员理论上能够、实际上也可以验证某个系统是否按照既定设计方案来运作。世界上不存在完全神秘的计算机程序算法。尽管某

〔35〕 关于深度学习的发展过程，参见集智俱乐部编：《深度学习原理与PyTorch实战》，人民邮电出版社2019年版，第3–4页。

〔36〕 参见周志华，《机器学习》，第6页。

〔37〕 Ajay Agrawal et al., *Prediction Machines: The Simple Economics of Artificial Intelligence* 105 (Harvard Business Review Press 2018).

〔38〕 Jeff Hawkins & Sandra Blakeslee, *On Intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines* 7 (Times Books 2004).

〔39〕 Ajay Agrawal et al., *supra* note 37, at 23, 76.

〔40〕 人脑和计算机的工作原理不完全相同。人类大脑皮层通过记忆、反馈、形成恒定表征（invariant representations）、建立并存储世界的模型对未来进行预测。智能机器则是利用已获得的数据生成人类尚未掌握的信息，或者说是通过填补缺失信息对未来作出预测。Id. at 29–30; Jeff Hawkins & Sandra Blakeslee, *supra* note 38, at 72–119.

〔41〕 “信息论之父”香农（C. Shannon）对什么是“信息”给出的解释是：不确定性的分解。这意味着，掌握越多的信息，就越能减少事物的不确定性。参见［美］C.R.劳：《统计与真理：怎样运用偶然性》，李竹渝等译，科学出版社2004年版，第106页。

〔42〕 参见康培和、徐奕奕：《计算思维：计算学科导论》，电子工业出版社2015年版，第5页。

些机器学习系统生成的结果难以在事前预测并且难以用传统方法进行解释,但采用机器学习这一决定本身就是设计系统时的人为选择。〔43〕

(2)算法是否“黑箱”?

有很多人认为算法的特征是黑箱(black box),应该说这是一种不完全正确的认识。〔44〕同算法的神秘性问题一样,算法并非都是黑箱,只有那些完全不公开、不透明的算法才称得上是“黑箱”。在很大程度上,算法黑箱是刻意保密的结果,是人为制造的,是权力的表达而非算法的本质。认清这一点至关重要:必须停止将算法神化为“黑箱”或者“魔盒”,因为这样我们正中了那些用不透明来为算法作掩护的人的下怀。〔45〕即使是典型的深度学习算法深层神经网络是一种难以解释的黑箱模型,但已有一些工作尝试改善神经网络的可解释性,主要途径是从神经网络中抽取易于理解的符号规则。〔46〕

不过,需要注意的是,尽管算法本质上并非黑箱,算法的黑箱化趋势却值得高度警惕。在数字化时代,同时存在两个相互背离的社会趋势:透明化和黑箱化。一方面,信息技术的发展使信息的获取、传递和交流更加便捷、迅速,任何物包括人及其活动都更容易被追踪、记录,社会结构的“粒度”更加精细,所有人都被高度解析,人们几无隐私可言,社会变得越发透明;另一方面,社会的黑箱趋势非常明显。许多数据巨头(私营部门)和政府机构借口“商业秘密”或“国家安全”竭力掩盖其使用的算法,以达到最大程度操控个人和社会的目的。透明化和黑箱化趋势并存,是数字化时代“去中心化”和“趋中心化”之外的又一个奇特悖论。

对于算法造成的黑箱趋势,美国马里兰大学法学院教授《黑箱社会》的作者帕斯奎尔(Frank Pasquale)曾提出一个耐人寻味的问题:如何遏制这种趋势并利用法律使黑箱社会变得透明?〔47〕申言之,我们必须从法律的角度思考如何揭开“盖子”,看看算法的葫芦里究竟卖的什么药?否则,黑箱社会一旦形成,或许真的会通过算法这种具有巨大破坏力的“数学杀伤性武器”(the weapons of math destruction)挖掘“算法陷阱”,最终建立一个由算法及其创造者统治世界的“算法帝国”。

回到现实中来。由于算法需要从求解方案的抽象描述转变为一个清晰的指令集合,并且这些指令需要用某种程序设计语言来表示并经编译程序编译为机器能够识别的二进制语言,在这个算法设计和编译的过程中,自动化决策系统不可避免地显现出一定的局限性。

(二)自动化决策系统(程序算法)的局限性

1.“智能机器”也会犯错

号称“智能”的机器不仅会犯错,而且有些错误非常离谱。例如,美国出现过多起国会议员、飞行员被错误列入恐怖分子“禁飞名单”而被拒绝登机的案例。再如,2010年5月6日,美国股市闪电崩盘,道琼斯指数在三分钟内下跌近千点。普遍使用的高频交易算法出错使市场陷入混乱,类似的情形曾多次发生。〔48〕

“智能机器”犯错的根源主要在于算法设计。在设计算法的过程中,有两个必不可少的步骤,一是对问题进行分析抽象出相应的数学模型;二是确定要使用的数据结构。数学模型是某个求解过程的抽象表示,没有模型能够囊括现实世界的所有复杂因素或者人类交流上的所有细微差别,有些信息会被遗漏。尤其是,模型的本质是简化,而简化导致的最严重问题是用线性代替非线性的情况,这是罪恶的普罗克拉斯提斯之床,也是一切问题的根源。生活中削足适履的典范莫过于简化非线性的事物,使之呈现

〔43〕 Joshua A. Kroll et al., *Accountable Algorithms*, 165 University of Pennsylvania Law Review 633, 637 (2017).

〔44〕 参见胡敏洁:《自动化行政的法律控制》,载《行政法学研究》2019年第2期,第61页;孙建丽:《算法自动化决策风险的法律规制研究》,载《法治研究》2019年第4期,第111页。

〔45〕 参见[德]库克里克:《微粒社会》,黄昆、夏柯译,中信出版社2018年版,第150–151页。

〔46〕 参见周志华,《机器学习》,第115页。关于算法的解释问题,后文作详细探讨。

〔47〕 Frank Pasquale, *supra* note 15, at 140.

〔48〕 参见斯坦纳,《算法帝国》,第2–5页。

出线性，这种简化扭曲了事实。[49]此外，如果算法是基于有缺陷、不完整甚或是错误的数据结构建立起来的，那么，不难想象，程序运行出现错误不可避免。

“智能机器”会犯错的另一原因，是由于那些采用深度学习的智能系统是从海量数据中提取特征以获得预测函数模型，在此过程中，需要不断地对各种模型参数进行调整，而这个调参过程实际上是不断试错，故而有人戏称其为“现代炼金术”，意指深度学习现有的理论基础尚不完全靠谱。虽然经训练成熟的深度学习系统在大多数情况下其输出结果惊人的准确，但它无法避免出现不知所云，甚至有时完全走偏的结果。特别是不需要人工标注数据的无监督学习算法，相较于监督学习，更像是让计算机在黑暗中自行摸索，其输出结果一定会存在“黑暗宇宙”，出错是必然的。正因为如此，采用深度学习算法的自动化决策系统虽然已经在商业领域先行一步，但它在行政领域的运用应该受到严格限制，以避免出现严重损害公民权利的情形。

2. 智能机器的“算法歧视”与人类的“自动化偏见”

除了算法设计的错误之外，算法当中还可能藏匿着某些隐而不彰的歧视和偏见。起初，人们相信，一台由算法驱动的计算机不受情感左右，对谁都会一视同仁（这也是自动化决策系统被广泛采用的重要原因之一）。然而，随着时间推移，人们开始对最初的设想抱有怀疑。有很多事例证明，某些算法当中潜藏着歧视和偏见。[50]

有学者指出，计算机系统中的偏见可分为三种：既存的偏见、技术性偏见和新生的偏见。既存的偏见（preexisting bias）是人类头脑中固有的偏见，其根源在于社会文化和社会制度；技术性偏见（technical bias）源于技术手段和设备的限制性因素；新生的偏见（emergent bias）则起因于应用场景的“时过境迁”，即原有的智能系统落后于社会发展。[51]

新近的研究发现，智能机器产生的歧视和偏见主要是第一种偏见，即社会文化和制度中既存的偏见。2017年4月14日，《科学》（Science）杂志刊发普林斯顿大学和英国巴斯大学一个联合研究小组的文章，证明人工智能系统中的歧视和偏见主要来源于人类，且人类社会现存的歧视和偏见会借助AI系统得到增强和放大。[52]不难理解，是人类设计了程序算法，他们固有的价值观和偏见可能会嵌入程序指令。所谓“垃圾进，垃圾出”（rubbish in, rubbish out），受到“污染”的程序算法不可能产生“出淤泥而不染的白莲花”。在现实生活中，常常有IT从业者以“技术中立”为由并以“菜刀”为例，为其非法行为进行辩护。这种抗辩不能成立。因为，信息技术与菜刀有根本的不同，每一种信息技术措施中都蕴含有人的意志，保障技术运行的算法和设计算法时预定的技术属性都具有某种倾向性，而菜刀没有。[53]

除了算法当中可能存在的歧视和偏见，还有一种因应用智能机器而产生的“自动化偏见”（automation bias）。由于自动化决策不仅比人类决策更为迅速，而且还有可能找到超越人类想象力的解题路径，由此让很多人误认为，自动化决策在任何情况下都可以提供比人类判断更好的结果。从本质上讲，自动化偏见来源于科学崇拜（科学主义）。对一般人而言，这种偏见似乎没什么大碍，但对于依靠、利用自动化决策结果的公务人员而言，自动化偏见十分有害，因为它客观上会强化“智能机器”的错误决策，使权利受到影响的当事人的处境更为不利。

3. 限缩裁量或放弃裁量

世界的纷繁复杂决定了不可能为所有的问题都预先设定好明确的羁束性规则，人们需要根据具体

---

〔49〕 参见［美］塔勒布：《反脆弱：从不确定性中获益》，雨珂译，中信出版社2014年版，第52、221页；［美］奥尼尔：《算法霸权：数学杀伤性武器的威胁》，马青玲译，中信出版社2018年版，第6–11页。

〔50〕 参见奥尼尔，《算法霸权》，第117页以下。Frank Pasquale, *supra* note 15, at 193.

〔51〕 Batya Friedman & Helen Nissenbaum, *Bias in Computer Systems*, 14 ACM Transaction on Info. Systems 330–347 (1996).

〔52〕 Aylin Caliskan et al., *Semantics Derived Automatically from Language Corpora Contain Human-like Biases*, 356 Science 183, 186 (2017).

〔53〕 美国科技史学家克兰兹伯格（Melvin Kranzberg）指出：“技术无所谓好坏，亦非中立。”参见［匈牙利］德士菲：《影子的社会学》，潘林峰译，载《学术评论》2015年第4期，第91页。

情境寻求有针对性的、灵活的解决方案。在行政领域,裁量权一向被认为是行政权的核心。没有裁量权,就难以实现个案的正当性。

但是,由于人性使然,裁量权极易发生滥用,因而必须对裁量权进行约束。这也是法治区别于人治的一个基本特征。〔54〕近年来,各国均大力发展电子政务,积极推进数字政府建设,广泛借助自动化决策系统作出行政决定,除了提高效率的考量之外,希望排除行政过程中的恣意是一个重要因素。

的确,自动化决策在很多领域胜过人类裁量。在需要前后一致超过裁量价值的情形下,实施良好的自动化决策比人类裁量更为可取;在与人类偏见有关的风险超过自动化偏见导致的风险之领域,自动化决策也更有吸引力。〔55〕简言之,当某一问题不需要根据特定情况进行裁量时,自动化决策具有人类决策无法比拟的优势。

然而,凡事皆有两面性,自动化决策的长处同时也是其短处。由于算法的确定性特征,在编程中不允许存在二义性,自动化决策系统难以像人类一样针对具体情境选择最适当的决定。这种在设计算法时就不得不限缩裁量甚至放弃裁量的做法,如果按照传统行政法学理论予以判断,明显属于“裁量怠惰”。

因此,自动化决策系统的适用范围具有局限性。波士顿大学法学教授希特伦(Danielle K. Citron)认为,最适宜运用标准(standards)而非规则(rules)予以处理的行政决定不能使用自动化决策;明确或者默示要求运用人类裁量的政策(policies)也不宜通过自动化决策来实施。〔56〕当然,希特伦教授的说法可以商榷。在其十多年前研究这一问题时,自动化决策多属于依靠规则推理的“专家系统”的运用。近年来,依靠大数据驱动的机器学习蓬勃兴起,自动化决策系统的适用范围有了新变化。不过,尽管如此,若要有效解决裁量问题,现在的自动化决策系统还面临艰巨的技术挑战。

4. 自动化决策的瞬时性危及程序保障

一般来说,计算机程序是由成千上百万条命令行组成的,这些命令行可以输出数亿条指示,最终的输出结果瞬间即告完成。自动化决策的时间往往以毫秒计,这也正是自动化决策的优势之所在。算法的价值全部体现在它的速度上。如果算法不能在毫秒或者微秒级内完成复杂任务,就不能称之为革新的力量。一个能给出正确输出但耗时很长的算法几乎没什么价值。算法设计的精妙之处,即在于如何多快好省地完成预定任务。说到底,智能既是一种实现复杂目标的能力,同时也是一种时间性能力。如果任务时间无限,即使是滴水也能穿石;要想办法用工具迅速穿石,这才称得上是智能。〔57〕

马克思曾言:“一切节约,归根到底是时间的节约。”无疑,自动化决策有助于大幅提高行政效率。然而,如前所述,自动化决策在带来便利和高效的同时,也引发了权利保障问题。智能机器中的算法有无错误?是否存在隐而不彰的歧视和偏见?有没有因算法放弃裁量而导致行政决定不合理?如果存在这些问题,当事人如何寻求救济?尤为重要的是,为保障当事人实体权利的实现并具有其自身内在价值的诸多程序保障如何通过算法设计得以体现?传统的正当程序在新技术条件下究竟还有没有可适用的空间?

三、算法权力的规制与技术性正当程序:人工智能时代正当程序的深化与发展

有学者指出:“法治的伟大成就之一是使主权者为其决策负责并赋予人民基本权利。……当新的算法决策者掌管了个人生活的方方面面,法律和正当程序在此领域若有缺席,我们就为这种不负责任的权

〔54〕 Bernard Schwartz, *supra* note 2, at 607.

〔55〕 Danielle K. Citron, *Technological Due Process*, 85 Washington University Law Review 1249, 1303 (2007).

〔56〕 *Id*. at 1304.

〔57〕 参见李彦宏等,《智能革命》,第245页;[美]泰格马克:《生命3.0:人工智能时代人类的进化与重生》,汪婕舒译,浙江教育出版社2018年版,第67页。

力中介人建立的一种新的封建秩序打开了方便之门。”〔58〕

有鉴于此，2007年，时任马里兰大学法学教授的希特伦提出了“技术性正当程序”（technological due process）的概念，强调通过优化计算机程序设计提高自动化决策系统的透明性和可问责性。〔59〕此前，德国著名行政法学家毛雷尔教授在1999年就已经提出，必须从法治国家的角度认识和规范“专门的技术程序”。〔60〕申言之，再复杂、高深的技术程序也不允许从法治国家的规范中逃逸，否则就会形成法治国家的“虫洞”，最终造成法治国只剩下一个“合法性的空壳”。

那么，能否以及如何构建“技术性正当程序”，从法律和技术两个维度规范算法权力呢？

必须指出，虽然适用的场景不同，但正当程序的价值理念诸如维护程序公平、保障人性尊严等，具有超越时空的特性，它们在新技术条件下也不会过时。因为人工智能技术的服务对象是人，只要人的主体性保持恒定，对人性尊严的尊重之要求亦不会改变。用计算机科学的术语来讲，正当程序的价值理念具有“鲁棒性”（robustness），即在受到持续扰动时仍保持原有的性能。

对正当程序理论稍有涉猎的人都知道：一旦确定要适用正当程序，接下来的问题就是，什么样的程序才是“正当”的？这一程序法的核心问题在人工智能领域则变换为：什么样的程序算法具有法律和技术双重意义上的正当性？如何通过程序算法的设计实现其正当性？

探索算法设计的正当性，仍然需要从正当程序的要求入手。关于正当程序的要求或判断标准，向来众说纷纭。例如，康奈尔大学法学教授萨默斯（R.S. Summers）曾提出正当程序的“十项程序价值”；佛罗里达州立大学哲学教授贝勒斯（M.D. Bayles）提出了正当程序的“八项程序利益”；威廉玛丽学院法学教授费奎尔（P.R. Verkuil）认为，不论程序的差别如何，一切对当事人不利的决定，必须包括四项最低限度的程序保障：（1）事先得到通知的权利；（2）口头或书面提出意见的机会；（3）决定必须说明理由；（4）作决定者没有偏见。〔61〕亦有中国学者根据德国思想家哈贝马斯的“交往行为理论”提出了判断正当程序的三项核心要素：排除偏见、听取意见和说明理由。〔62〕沿着这一简单明了且容易作形式化审查与判断的思路进一步推论：这三项要素应当在自动化决策系统的算法设计中继续适用，并得以彰显和体现。

（一）排除偏见：算法的公开、透明和程序一致性

我们知道，在法律程序中，排除偏见要求裁判者不能做自己案件的法官，或者说，裁判者自身不能与案件存在利益上的关联。而在自动化决策系统中，智能机器谈不上自己的利益，排除偏见的要求是排除算法歧视。排除算法歧视的技术，一是借助于算法的公开、透明，二是确保程序的一致性。

1. 算法的公开、透明

传统上，法官在作出裁判之前需要进行充分的说理和论证，说理意见以判决书的形式供公众审阅。行政决定的作出过程也是如此。但是，行政自动化决策系统并不如此运作，普通人无法理解算法的原理和工作机制，自动化决策系统也常常是在算法的“黑箱”中作出决策，不透明性问题由此产生。

显然，我们不可能和“黑箱”建立信任关系。算法黑箱不仅容易滋生腐败、造成侵权和伤害，而且，透明性缺失反过来也会动摇、瓦解公众的信赖，导致自动化决策系统的应用与实施停滞不前。例如，2019年9月28日，杭州某楼盘在购房摇号时出现大面积集中连号。10月12日，负责此次摇号公证的杭州市国立公证处发布消息称：“浙江千麦司法鉴定中心出具的鉴定意见显示：9月28日的摇号中，第一轮摇号

〔58〕 Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 Washington Law Review 1, 19 (2014).

〔59〕 Danielle K. Citron, *supra* note 55, at 1249–1313.

〔60〕 毛雷尔，《行政法学总论》，第442–443页。

〔61〕 *See* Robert S. Summers, *Evaluating and Improving Legal Process: A Plea for "Process Values"*, 60 Cornell Law Review 1–52 (1974); Paul R. Verkuil, *A Study of Informal Adjudication Procedures*, 43 University of Chicago Law Review 739–796 (1976).［美］贝勒斯：《程序正义：向个人的分配》，邓海平译，高等教育出版社2005年版，第155页。

〔62〕 参见刘东亮：《什么是正当法律程序》，载《中国法学》2010年第4期，第76页。

的数据未发现异常,第二轮摇号存在部分数据未被赋予随机值的异常情况。”尽管事后进行了重新摇号,该公证处也向社会表示了道歉,但该事件造成的恶劣影响可想而知。〔63〕

加强算法的公开和透明势在必行。多国发布的政策均提到AI决策的透明度对于监管的重要性。2016年9月,英国下议院科技委员会发布的《机器人与人工智能》,强调了决策系统透明化对人工智能安全和管控的作用。〔64〕在美国, 100多年前,大法官布兰代斯(Louis Brandeis)指出:“阳光是最好的消毒剂。”至今,这句话仍然被频繁引用。因而,美国的法律和政策非常注重透明度,有时甚至将监督等同于透明度。〔65〕2019年6月,美国国家科技委员会发布《人工智能研发战略计划》(2019年更新版),强调通过优化设计提高AI系统的公平、透明和可问责性。〔66〕几乎同一时间,中国政府发布《新一代人工智能治理原则》,要求“人工智能系统应不断提升透明性”。〔67〕

那么,如何实现算法的公开、透明呢?解决这一问题,首先需要弄清楚算法缺乏透明的原因。有学者指出,产生不透明性的情况有三种:一是前文所述因刻意保密产生的不透明;二是因技术无知(technical illiteracy)产生的不透明;三是由于机器学习算法的特征和有效运用机器学习算法所需要的规模与程度产生的不透明。其中,第二种不透明源于代码编写和算法设计在目前仍然属于专家技能,第三种不透明源于机器学习高维特征的数学优选方法和人类尺度的推理(human scale reasoning)及语义学解释风格之间的不相匹配(简言之,人类目前的认知能力尚难完全揭示机器学习的逻辑),并且,后两种形式的不透明常常是混合在一起的。〔68〕

这三种形式的不透明性可使用不同的工具和方法来处理,包括立法性的、组织性的或者编程性的、技术性的工具和方法。其中最重要的是首先识别出不透明性的具体形式,然后确定在特定的情形下应当采用哪种技术性和/或非技术性的解决方案防止损害的发生。〔69〕

对于第一种形式的不透明,即因保护商业秘密或国家秘密而造成的不透明,有人担心,如果强制公开可能引发严重的“并发症”。因为,自动化决策系统想要在“猫鼠游戏”的博弈中具有有效性,就得刻意保密以避免对手甚至是犯罪分子的“反向学习”(adversarial learning),因此,大多数算法向社会公开披露根本不可能。〔70〕对此,帕斯奎尔认为,即使这种不透明性也是一种“可矫正的不可理解性”(remediable incomprehensibility)。他主张借助于独立的、可信任的算法审计人员使算法受到审查,同时又能维护真正的商业秘密和公共利益。〔71〕试想,如果连政府都不公开“黑箱算法”,受到众多私营机构联合操控而正在形成中的“黑箱社会”确有变成“黑社会”的危险。

当然,以必要的规制手段(例如立法)强制要求公开自动化决策系统的算法源代码或者相关的数据库,并不意味着无选择、一刀切地全部公开披露。任何旨在解决黑箱问题的透明性方案都应该至少明确三个问题:披露多少(how much)、向谁披露(to whom)和披露的速度(how fast)?这三个问题都不是无限制的。即使是解密者阿桑奇(Julian Assange)和斯诺登(Edward Snowden)也对选择性披露的做法予以认同,他们在披露时都对可靠的信息源进行了过滤。帕斯奎尔称这种要求为“优质的透明度”

〔63〕 参见陈洋根:《受“连号”事件影响杭州一楼盘重新摇号》,载《浙江法制报》2019年10月29日,第7版。

〔64〕 参见腾讯研究院等,《人工智能》,第195页。

〔65〕 Joshua A. Kroll et al., *supra* note 43, at 705.

〔66〕 National Science & Technology Council, *The National Artificial Intelligence Research and Development Strategic Plan (2019 Update)*, June 2019.

〔67〕 参见中国国家新一代人工智能治理专业委员会:《新一代人工智能治理原则:发展负责任的人工智能》, 2019年6月17日。

〔68〕 Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, 3 Social Sciences Electronic Publishing 1–12 (2015).

〔69〕 *Id*. at 3.

〔70〕 C. Sandvig et al., *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*, in *Annual Meeting of the International Communication Association*, Seattle, WA, 1–23 (2014).

〔71〕 Frank Pasquale, *supra* note 15, at 7, 141, 218.

(qualified transparency),即为了尊重某一信息涉及的各方利益,必须对披露的范围进行限制。[72]

对于第二种和第三种情形的不透明,即由算法的复杂性导致的不透明性,无论主动公开还是强制公开,其作用都非常有限。因为,宛若“天书”一样的代码,在大多数人眼里,都只是一些杂乱无章的“乱码”而已。在此情况下,必须要由专业人士通过算法审计 (algorithm audit) 判断是否存在算法歧视,包括是否存在数据的歧视性分类与标记。算法审计的标准是要求其具有“程序一致性”。

2. 程序一致性

程序一致性 (procedural regularity) 是正当程序的一项核心理念,其含义是:规则必须一体适用,不能单为某个人选择不同的程序。[73] 易言之,同样处境的人应该同样对待,不应考虑敏感的身份特征。

要使自动化决策系统中的算法具有程序一致性,核心是确保规则或政策的普遍适用性 (这也是法律的基本特征之一)。此外,从结果来看,还应确保决策的可重复性和可再现性。申言之,自动化决策系统需要确保:(1) 每一项决策都运用相同的规则或政策;(2) 在知道特定的决策对象之前,决策所依据的规则或政策应当明确、具体,减少规则或政策选择因人而异且有不利于特定个人的可能性 (用中国的俗话来说,避免“看人下菜碟”);(3) 在相同条件下,每项决策都是可重复、可再现的;(4) 如果决策要求随机选择的输入,这些输入超越任何利害关系人的控制。

上述考虑需要纳入自动化决策系统的初始设计,并采取相应的技术手段保证其可以实现。这就要求开发有关的技术措施,比如预设某些屏蔽敏感特征的规则。一个具体的例子是只能为了某些目的使用某些信息。例如,在一个职位申请人筛选系统中,系统允许输入申请人的性别,但性别输入只能用作信息统计的数据,而不能作为职位筛选之目的。[74]

算法设计是否做到了程序一致性,需要对之进行评估和检测。通常,计算机科学家使用两种方法对算法进行评估:(1) 静态分析 (static analysis),即不运行程序而只检查源代码;(2) 动态分析 (dynamic analysis),即运行程序评估特定输入后的输出,或者评估程序运行时的状态。[75] 其中,动态分析既可以运用“黑箱测试”,即只考虑系统或组件的输入、输出,也可以运用“白箱测试”,即考虑系统的内部结构以设计测试案例。[76] 在测试评估时,审计人员可借助专门的软件测试工具,包括自动化软件测试工具和测试管理工具。一个好的软件测试工具和测试管理工具结合起来使用,可以大大提高算法测试的效率。[77]

程序一致性能够有效排除算法歧视,保证一以贯之地适用相关规则或政策。不过,程序一致性仅仅是确保程序公平的一项原则,决策过程本身是否合理,尤其是决策结果是否基于可靠的推理 (sound reasoning) 是独立于程序公平的另一个问题。这一问题需要结合算法的可解释性进行判定。

(二) 说明理由:算法的可解释性

普林斯顿大学信息技术政策研究中心克鲁尔教授 (Joshua A. Kroll) 指出,管理自动化决策的首要目标,是让决策过程的监督者 (无论是政府官员、企业高管还是社会公众) 知道计算机系统是如何做出决策的,或者至少知道是基于哪些规则做出决策的,即使该规则不完全公开。[78] 这里所说的“知道”不同于传统法律所讲的“知道或者应当知道”,其责任应由算法的设计者来承担 (举证责任倒置)。易言之,“知道”的责任通过算法的可解释性 (interpretability) 来实现。

---

〔72〕 *Id*. at 142.

〔73〕 Shaughnessy v. Mezei, 345 U.S. 206, 224 (1953).

〔74〕 Joshua A. Kroll et al., *supra* note 43, at 657.

〔75〕 Gary McGraw, *Software Security: Building Security In* 119–121 (Addison Wesley Professional 2006).

〔76〕 *Id.* at 200; Joshua A. Kroll et al., *supra* note 43, at 650–651.

〔77〕 由于算法的评估和测试方法涉及到复杂的计算机技术问题,本文不再详细展开。除前述Gary McGraw的著作和Joshua A. Kroll等人的论文外,有兴趣的读者可参阅“软件测试”的相关文献。

〔78〕 Joshua A. Kroll et al., *supra* note 43, at 656.

1.为什么强调算法的可解释性

如前所述,算法的公开、透明非常重要。但算法的公开、透明只能解决部分问题,或者仅仅为问题的解决创造条件而已。说到底,算法公开、透明,最直截了当的方法莫过于公开源代码。然而,面对一行行、一页页、一堆堆“蝌蚪文”一样的代码,非专业人士肯定是一头雾水,只能知难而退。即使是在面对专业人士时,被监督对象也总能找到应对的办法。例如,当某公司需要提供相关信息应对调查时,它可能给出3000多万页的文件资料,让调查人员在调查时如同海底捞针、水中捉月〔79〕——透明性可能引发复杂性,而复杂性一样会妨碍理解。

事实上,算法的公开、透明对于实现算法问责既不充分,也不总是必要的。透明性本身不是目的,它只是通向可理解性(intelligibility)的一个阶梯。〔80〕相对于透明性,可理解性才是目的。

可理解性和可解释性是一体的两面。如果某个自动化决策系统中的算法完全不可理解、不可解释,毫无疑问,这种算法蕴藏着巨大的风险。构建一个人类完全不能理解的系统(它有很大可能也无法被控制),相当于人类亲手制造了一个“潘多拉魔盒”,这肯定是我们无法承受的。在算法作出任何与人类的生命、自由或财产息息相关的决定时,人类有权知晓它背后的逻辑和理由,否则人类就被粗暴对待而降级成“奴隶”或“客体”。

可解释性是人类与自动化决策系统之间的接口(interface)。〔81〕国际电气电子工程师协会(IEEE)发布的《人工智能设计的伦理准则》要求自动化系统为其决策提供明确的理由,该准则第二版再次重申了自动化决策系统应具有“可解释性”。〔82〕2018年5月生效的欧盟《通用数据保护条例》(GDPR)在序言部分规定,当数据主体对自动化决策有异议时,有权获得人工干预、表达意见、获得对决策的解释,并对决策提出质疑。该条例第13条规定,在存在自动化决策的情况下,数据主体有权要求数据控制者提供决策的相关逻辑及此类处理对于数据主体的预期后果的实质性信息。〔83〕

2.如何对算法进行解释

算法的解释方法因算法模型的不同而有较大差异。影响算法可解释性的关键因素,除了算法的透明度(白箱模型或黑箱模型),还有算法结构的复杂度。

一般来说,白箱模型很容易作出解释。诸如决策树、规则列表、线性回归等算法透明、结构简单的模型,都是人类可理解且易解释的。因为这类模型皆属于或者可转化为“if-then”形式的决策规则,其中的逻辑推理很容易为人类所把握。当然,必须承认,如果使用数百个甚至上千个特征构建一个非常庞大的决策树,那么,这种算法也会因为复杂性而削弱自身的可解释性。

典型的黑箱模型,如深度神经网络(DNN)、随机森林(RF)和梯度增强机器(GBM)的可解释性较弱。对于这类算法,只能借助于其他方法提供一个粗粒度的解释。但需要指出,这类算法并非完全不可解释,不存在绝对的黑箱。以DNN为例,在训练样本数据时所进行的人工标注,是可解释的;虽然深度学习的参数调整高度依赖人的经验,它也并非完全不可解释。事实上,深度学习的参数调整以试错法为基础,这与自然科学研究中普遍运用的方法在本质上并无不同。在追求目标的过程中不断试错和消除误差,根据试验对象作出的应答与反馈,找到实现目标的最佳途径,也是我们人类在世界上运用最普遍的方法。因此,对于黑箱模型,要想方设法将其转化为“灰箱”,或者说,要尽可能将其中的“黑暗宇宙”压缩到最小空间。办法总比困难多。

在具体的解释方法上,需要根据具体情况灵活运用或者综合运用下列方法:

(1)预建模的解释和建模后的解释。预建模的解释是指在建模之前就考虑使用的数据特征,

〔79〕 Frank Pasquale, *supra* note 15, at 6–8.

〔80〕 *Id.* at 8.

〔81〕 Riccardo Guidotti et al., *A Survey of Methods for Explaining Black Box Models*, 51 ACM Computing Surveys 93:5 (2018).

〔82〕 IEEE, *Ethically Aligned Design*, Version 1 & 2.

〔83〕 *General Data Protection Regulation*, Recital (71) & art. 13(2)(f).

使算法本身内置可解释性，由于算法的可解释性发生在模型建立之前，因而也称事前解释（ante-hoc interpretation）。事前解释通常针对结构简单、易于理解的算法。而建模后的解释（post-hoc interpretation）针对透明度较差、结构复杂的算法，需要开发相关的解释性技术甚至建立解释模型来解释算法的工作机制和决策依据。其中，可解释性技术作为一种细粒度分析和解释模型的有效手段，可用于分析、调试模型的错误决策行为，诊断模型中存在的缺陷，并为修复模型缺陷提供有力支撑。〔84〕

（2）全局解释与局部解释。全局解释是从整体上解释算法背后的复杂逻辑及其内部的工作机制。典型的全局解释方法包括规则提取、模型蒸馏、激活最大化解释等。在实践中，由于算法的不透明性、结构复杂性及应用场景的多元性，提供全局解释通常比提供局部解释更为困难，因而局部解释相对于全局解释更为常见。经典的局部解释方法包括敏感性分析解释、局部近似解释、梯度反向传播解释、特征反演解释以及类激活映射解释等。〔85〕

需要指出，近几年来，如何改善深度学习的可解释性较差问题，已成为人工智能研究的一个热点，并取得了明显的进展。例如，西安交通大学数学与统计学院徐宗本教授和孙剑教授2018年1月在《国家科学评论》（National Science Review）联合发表的论文《模型驱动的深度学习》，探索结合模型驱动和数据驱动方法，解决现有的启发式方法存在的普遍限制，使深度学习的可设计性和可预测性在一定程度上成为可能。〔86〕再如，2018年3月，谷歌大脑团队发布了一项重要的研究成果《可解释性的构成组件》（The Building Blocks of Interpretability），探讨了如何将机器学习领域正在发展中的几种强有力的可解释性技术，如特征可视化（feature visualization）、归因法（attribution）和降维法（dimensionality reduction）组合起来探索神经网络如何决策。这种组合方法让我们“站在神经网络内部”，看到神经网络在某一具体时刻如何工作及如何影响最终输出。也就是说，这种组合方法能够显示神经网络在识别图像时发现了什么并解释其是如何发展出某种理解的，同时在所需要的信息量上保持“人类尺度”，这使得可解释性有望成为在塑造有意义的人类监管制度和建设公平、安全、均衡的人工智能系统时的强大工具。〔87〕

当然，深度学习的可解释性虽已取得了一定的进展，但目前仍处于初级阶段（点上有推进，面上待突破），还有大量问题尚待解决。而且，不同学者解决问题的角度不同，对可解释性赋予的含义不同，提出的解释方法也各有侧重。〔88〕

3. 算法可解释性的保证：提供审计跟踪记录

与算法解释密切相关的一个重要问题是自动化决策系统应当提供不可篡改的审计跟踪记录（audit trails）。这种类似飞机“黑匣子”的记录不仅使行政机关可以向当事人解释自动化决策所依据的事实与规则，更重要的是保证了决策的可追溯性（traceability）。可追溯性让人类监管机构不仅能够理解智能机器的决策过程并在以后作出必要的修正，而且能够在特定的调查和法律行动中发挥其应有的作用。特别是，审计跟踪记录有助于纠正听证官员持有的自动化决策系统一贯正确的推定（即前文所述“自动化偏见”）。有鉴于此，IEEE《人工智能设计的伦理准则》多处要求自动化决策系统提供审计跟踪记录。澳大利亚等国已经通过行政规章规定自动化（辅助）决策系统应当提供审计跟踪记录。〔89〕

（三）听取意见：允许质疑、事后听证、专业审计、及时纠错

1. 对自动化决策应当允许提出质疑和挑战

同人类决策一样，由智能机器作出的自动化决策出现错误难以完全避免。在目前的技术条件下，人

〔84〕 参见纪守领等：《机器学习模型可解释性方法、应用与安全研究综述》，载《计算机研究与发展》2019年第10期，第2085-2086页。

〔85〕 参见纪守领等，《机器学习模型可解释性方法、应用与安全研究综述》，第2071-2096页。

〔86〕 *See* Zongben Xu & Jian Sun, *Model-Drive Deep Learning*, 5 National Science Review 22-24 (2018).

〔87〕 参见Chris Olah et al., *The Building Blocks of Interpretability*, 资料来源：https://distill.pub/2018/building-blocks/，访问日期：2020年3月21日。

〔88〕 参见吴飞、廖彬兵、韩亚洪：《深度学习的可解释性》，载《航空兵器》2019年第1期，第43-44页。

〔89〕 Australian Administrative Review Council, *Automated Assistance in Administrative Decision Making*, Issues Paper 31 (2003).

工智能有时表现为让人啼笑皆非的“人工智障”。在安全性方面,对于采用机器学习算法的系统,别有用心者利用对抗样本(adversarial examples)攻击技术,在输入样本中添加精心构造的、人眼不可察觉的扰动就可以轻松地让决策系统出错。不仅如此,具有深度学习能力的智能系统一次很小的失误,其错误可能会在后续决策中得到增强。归根到底,算法决策是用过去的数据预测未来,而过去的错误或者蕴含的歧视可能会在算法中得到巩固并在未来被放大,因为错误输入形成的错误输出会作为负反馈进一步加深错误。因此,对于运用自动化决策系统作出的决定,必须允许提出质疑和挑战。也就是说,对人工智能决策结果一定正确的迷信需要破除。

2.自动化决策的适用场景只能采取事后听证

在传统行政法领域,尽管听证通常是要求某种形式的听证,而不要求固定形式的听证,采用事前听证还是事后听证应当根据具体情形来确定,但由于自然公正或者正当程序的原初含义是受到政府决定影响的当事人有权要求“事先通知(advance notice)和由中立的裁判者主持的公平听证(fair hearing)”,〔90〕因而,正当程序的传统模式仍然要求在不利决定作出之前进行听证。简言之,事先听证是行政听证的一般情形。

不过,在适用自动化决策的场景中,采取事后听证相当确定,不可能在毫秒甚至微秒级内先行听证。事后听证才是自动化决策场景的一般模式,这与行政听证的普通情形有所不同。这也意味着:自动化决策不能完全免除听证,但听证的形式可以变通。

3.在专业审计人员的协助下审查算法并及时纠错

在对自动化决策系统进行监督或者寻求救济的过程中,无论采取复议还是诉讼的方式,专业算法审计人员的协助都必不可少。前述杭州楼盘摇号案由司法鉴定中心出具鉴定意见,就是一个具体的例证。事实上,聘请独立的第三方算法审计机构对算法进行定期检测,也是维护自动化决策系统正常运作不可或缺的一环。〔91〕

不难理解,及时纠错对维护自动化决策系统的正常运作十分必要。前文指出,“代码即法律”,具有深度学习能力的自动化系统犯下的小小失误有可能会在后续决策中得到增强,存在错误的算法及其代码如果不及时修正,很快就会变成令人不能容忍的“恶法”。从技术的角度和软件开发的全过程来看,软件开发各个阶段之间的关系通常也不是顺序的、线性的,而是带有反馈的、不断纠错的迭代过程,这种纠错的迭代过程延续到软件的运行与维护过程之中。〔92〕

四、结语:人工智能的未来取决于人类如何行动

人工智能的未来会是什么样,没有人知道其确切答案。人工智能的进步会不会最终因为某些不可跨越的桎梏而停滞下来?人类最终能否成功地创造出人类水平的通用人工智能?对于这些问题,截至目前,世界顶级人工智能专家各执一词。没有人保证我们能够建造出人类水平的通用人工智能,但是也没有绝对确定的理由说,我们永远无法建造出来。〔93〕

尽管如此,可以肯定的是,随着人工智能技术的快速发展和广泛应用,如何应对其所带来的一系列深刻变化已成为人类命运共同体必须面对的一个全球性问题。同样可以确定的是,人工智能未来的走向如何,最重要的是人类如何行动。〔94〕因为人类是人工智能的缔造者,人类对其结果具有很大的影响力,

〔90〕 John Alder, *Constitutional and Administrative Law* 394 (Palgrave Macmillan 2013).

〔91〕 Joshua A. Kroll et al., *supra* note 43, at 641.

〔92〕 参见周玉萍,《信息技术基础》,第242页以下。

〔93〕 泰格马克,《生命3.0》,第172–175页。

〔94〕 李开复:《AI·未来》,浙江人民出版社2018年版,第354页。

这个影响是我们在创造人工智能的过程中潜移默化地加诸其上的。〔95〕人工智能的早期奠基人之一西蒙（H.A. Simon）在被问到计算机将如何继续塑造世界时也回答说:“实质上,虽然计算机将表现出巨大的力量,但是,如何接受和使用这种力量将依然取决于人。”〔96〕

因此,人工智能的未来在很大程度上取决于人类如何对其进行设计。进一步说,这涉及两个重要问题:如何设计和由谁来设计。

如何设计? 当代著名工业设计管理大师布鲁斯（Gordon Bruce）曾指出:“设计的本质与技能无关,而是一种生活态度。设计者需要重新设计自身思维方式,带着责任感、敬业精神和对人与环境的尊重去做设计。”〔97〕申言之,只有将人的主体性和对人性尊严的尊重放在首位,正当程序的理念才可能在人工智能算法的设计过程中得到贯彻,“以人为本”为核心价值的阿西莫夫三原则（机器人三原则）才不至于流于空想层面。

谁来设计? 在人工智能时代,不仅必须要由法律学者和计算机专家联手才能塑造出正当程序的轮廓,而且,还需要心理学家、社会学家、人类学家及其他一切与“人”的科学有关的专家的参与,才能设计出更懂得人、更理解人、更尊重人、也更能服务于人类福祉的人工智能系统。因为,对人工智能的研究,关系到人的哲学基本问题,涉及人文主义的核心关切,需要我们重新认识人类自身,也需要人文科学、社会科学和自然科学的综合研究共同承担。〔98〕从这个意义上说,本文从“技术性正当程序”视角的考察,只是一种初步的探路者作用,对人工智能技术特别是对其程序算法的探索,还有许多未竟的工作。

Technological Due Process: The Double Variation of Procedural Law & Programming Algorithms in the Age of AI

Liu Dongliang

Abstract: The wide application of artificial intelligence (AI) has been reshaping the operation mode of government, which leads to the changes of power form and structure in modern society. Thus, algorithmic power quietly springs up. At the same time, traditional administrative procedures are not feasible and application on the scene of AI. For instance, procedural safeguard measures such as hearing, giving reasons and so on can hardly work for automatic machine which make their decisions instantly. Under such circumstances, the focus attention of administrative law should be expanded from traditional administrative procedure to AI programming algorithms. The regulation and supervision of programming algorithms require construction of technological due process at the beginning of algorithm design. That is to say, algorithm design should be in a transparent and interpretable way and conform to procedural regularity. AI can provide related meaningful information about the logic of their decisions. Interested parties have the right to challenge the decisions; the review body audits programming algorithms with the help of independent third auditor; the error should be revised instantly. Technological due process requires that administrative law keep pace with the times and technical development.

Keywords: artificial intelligence; algorithmic power; algorithm design; technological due process

（责任编辑:刘 馨）

〔95〕 泰格马克,《生命3.0》,第213页。

〔96〕 参见卢奇、科佩克,《人工智能》,第210页。

〔97〕 Gordon Bruce, *Redesigning the Mind = Becoming More Mindful about Design*, 2019年9月9日下午3时在西安交通大学科学馆的演讲。

〔98〕 参见韩水法:《人工智能时代的人文主义》,载《中国社会科学》2019年第6期,第25、42页。