·第五届（2022）全国青年理论创新奖征文选登·

“无代表，不算法”：算法统治的政治代表性问题

王中原

【内容摘要】 算法统治（algocracy）是公共部门利用算法技术创造的一种新型统治秩序和治理形态。政务算法化正席卷各类政体，不断放大国家的规制性权力、强制性权力和分配性权力，重塑权利与权力、自由与秩序、公平与效率之间的平衡关系。当前，算法统治在算法输入、运算和输出环节存在授权、问责、回应方面的实质性代表性问题，造成公民声音的不在场性和公民权益的损失风险。同时，算法统治倾向于不断沉积、固化、强化乃至结构化特定群体的弱势地位，排斥少数群体的平等出场和权益实现。建构“算法代表性”有助于约束“算法利维坦”的权力边界和运行方式，保障公民特别是弱势群体的算法权益，提升算法统治的合法性，同时拓展新技术环境下的代表理论。

【关 键 词】 算法统治　算法代表性　算主社会　算法政治　政务算法化

【作　　者】 王中原，复旦大学社会科学高等研究院副教授、复旦大学复杂决策分析中心研究员。（上海 200433）

【基金项目】 国家社科基金青年项目“算法治理视域下政府数字化转型的实践困境及其破解机制研究”（21CZZ039）

导言

算法世界如何处理政治代表问题？算法驱动的决策系统越来越广泛地决定着我们的公共生活，[①] 每个人都直接或间接地卷入新型的“算法统治”或者“算主”体制（algocracy）。[②]政务算法化开始席卷各类政体，算法决定社会福利和公共物品分配，开展执法监督和治安防控，参与应急管理和疫情防治，实施边境管控和难民治理，执行计算宣传和信息投喂，承担社会信用评级和犯罪风险预测，算法甚至可以控制智能武器裁决生命权利。[③]政务领域的数字化转型正在重塑公

共权力的组织和行动模式，改变政府与公民互动的场景和界面，产生新技术环境下的复杂政治关系。由此，算法世界的授权、问责和回应等政治代表形态不同于传统政治世界，且处在早期孕育和型构阶段。探究算法统治下的新型政治代表关系和建构算法代表性，有助于改善公共事务中的算法治理，并推动新技术环境下的政治理论创新。

算法统治如何安顿权利与权力、自由与秩序、公平与效率的关系？算法部署在哪些公共场景中？谁有权决定引入算法系统和使用公民数据？算法过程如何构建模型、挖掘数据和自动决策？算法结果产生哪些政治社会影响？脆弱群体的权益如何在“算主社会”得到救济和保障？这些前沿问题本质上都是政治问题，然而一直缺少基于政治学理论的解释框架。当前，相关讨论要么着眼于公共管理和应用技术层级，要么聚焦在哲学辨析和伦理反思层面。现有研究较多关注算法的积极治理效能，即算法科技的数字赋权和数字赋能；或者关注算法异化，包括算法操控、剥削和杀熟，以及算法的不正义和伦理风险。如何从政治学角度重新审视算法治理问题？本研究尝试接续政治代表理论的前沿发展，探讨和反思算法统治下的政治代表关系，同时借助算法场景拓展传统的政治代表理论。

政治代表（political representation）旨在让不在场的权益和声音获得再次出场（re-present）的机会，是政治现代化的核心内容。代表不等同于“代议”，政治代表是在特定场域中建立起的权利与权力关系模式。该模式既要求被代表的诉求和利益传导进政治统治和决策过程当中，影响代表者的行为；又要求代表者接受被代表者的问责，并回应被代表者的诉求。长久以来，选举被认为是实现政治代表的经典途径。然而，在人类政治生活的众多场域，政治代表关系并非且较难借助选举达成。例如，在国际组织的运行当中，在社会运动的组织当中，在网络空间的意见表达当中，并未通过选举建立正式代表关系，但其间不乏选举之外的政治代表性。由此，近年来学界形成了“非选举型政治代表”的前沿思潮和理论转向，旨在探索非选举非代议场景下和复杂公共生活中的政治代表现象。

公共生活的数字化转型和政务活动的算法化，为我们思考和建构政治代表关系提供了全新的场景。算法统治将产生哪些政治代表性困境？当算法创造了新型的治理结构和统治秩序，而该统治无法按照传统的选举模式开展授权、问责和回应时，我们应该如何设计和运行新的代表关系模式，从而让算法服务于民众福祉且将算法“关在笼子里”？算法时代，我们需要重新审视政治代表问题，在自上而下的监管模式之外，探索自下而上的代表模式。

本研究试图从政治学理论的角度考察算法治理问题，聚焦政务算法场景下的政治代表困境及其应对策略，尝试将公共管理问题政治学化。首先，本文将在政治代表前沿理论的基础上尝试建构“算法代表性”（algorithmic representation）的新概念和分析框架。接着，从授权、问责和回应三个实质性代表维度考察当前算法统治的代表性困境，集中分析算法过程的输入、运算和输出环节存在的代表性问题及其生成逻辑和技术机理，并通过案例分析揭示算法的代表性危机可能导致的政治社会影响。最后，结合算法治理实践，探讨如何在算法统治下重新“创造在场”（creating the presence），迈向“算法代表性”，进而丰富新技术环境下的代表理论。

文献综述和理论基础

算法统治是公共部门利用算法技术创造的一种新型统治秩序和治理形态，算法不断放大国

家的规制性权力（regulatory power）、强制性权力（coercive power）和分配性权力（distributive power），由此形成新的权力与权利关系场域。在“算主”体制下，重新理解、评估和建构政治代表性将成为理论研究和实践探索的前沿。如果说霍布斯意义上的“利维坦”是自然状态中的人出于对战争和死亡的畏惧，在理性指引下订立契约，放弃个人部分自然权利并将之委托出去组成国家，同时国家承担保护个人安全和基本生存的责任，由此形成原初形态的代表关系，那么，面对“算法利维坦”（Algorithmic Leviathan），④我们应该如何思考和建构新型的政治代表关系？这首先需要回到政治代表理论脉络当中，并接续代表理论的前沿浪潮。

政治代表是一个内涵丰富、结构复杂和处在不断丰富发展中的政治学理论体系。从词源 representation 来说，政治代表是指某种利益或声音的政治再现，即让不在场者的权益或声音获得出场的机会（make present again）。⑤在不同政治场域中，代表关系建构和运行呈现出不同的模式。围绕如何实现政治代表性，也涌现出不同的理论学说，为算法场域的代表性分析提供了理论基础。

政治代表理论认为代表关系的形成依赖三个核心机制：授权（authorization）、问责（accountability）和回应（responsiveness）。具体而言，被代表者首先需要表达同意和授权，通过特定程序将代表内容委托给代表者；接着，获得授权的代表者需要回应被代表的利益和诉求，在政治过程中再现被代表者的声音；此外，代表者在行使授权时须负担起相应的责任，接受被代表者的监督和问责。⑥传统的政治代表理论强调授权、问责和回应的过程必须是正式的，即经由制度化的选举程序才具合法性。基于选举的政治代表理论统摄了政治学半个多世纪，选举型政治代表制之所以成为主流，是因为选举提供了制度化、操作化、周期性和合法性的授权和问责机制，被认为能够有效约束代表者。

然而，现实政治生活中并非所有的代表关系都依赖正式的选举程序。近年来，新型的政治代表场域不断涌现，⑦包括国际组织、社会运动、公益行动、智库机构、倡议团体和网络意见领袖等，这些领域的代表关系并不依赖选举程序，但其代表性的强度和质量并不亚于民选代表。因此，越来越多的政治理论家意识到有必要将代表问题与选举问题脱耦，探索选举之外的授权、问责和回应机制，进而形成“非选举型代表”的理论前沿。

第一，代表理论出现建构主义转向（constructivist turn）。⑧传统的选举型政治代表预设了民众已经形成自己的偏好，选举只是将个体偏好聚合成群体偏好。但是，现实中选民偏好并非总是先存的（preformed），代表者也不仅仅是“传声筒”，偏好完全可以被建构。例如，萨沃德提出“代表宣言”学说，⑨强调代表者的任务是提出响应特定对象的某种陈述，代表政治不仅意味着应对性的行动，还包括前瞻性的引领，该偏好建构过程无须经历选举程序。第二，代表内容转向多元话语。德雷泽克和尼迈亚提出“话语性代表”的概念，⑩不同话语的代表者参与正式或非正式的协商，并在协商中不断校正话语体系以达成共识。⑪与此类似，乌彼莱特主张将代表视作“倡议”（representation as advocacy），⑫即通过倡议反映社会利益的多元性和异质性，而非通过投票聚合偏好掩盖差异性和不平等。第三，代表的动力来源由外转内。曼斯布里奇指出经典的政治代表理论太过强调“惩罚模型”，即代表者面对败选威胁才被迫承担责任，该模型忽略了自我驱动型代表者。⑬代表可以主动担当，⑭自行宣称代表特定群体的权益，并作为积极能动者参与公共事务的治理。第四，代表者从精英转向公民。在网络空间、社会运动、自治团体等场域并不存在职业化的政治代表，也没有选举式“委托–代理”程序，代表者和被代表者的身份通常是交叉和流动的。沃伦据此提出了“公民代表”路径，主张公民通过非选举途径成为社群代表，并依靠培养判断能力和

协商素养提升代表质量。[15]

上述政治代表理论的前沿发展呈现出“非选举型代表”的理论转向，这些学说从代表者、被代表者、代表内容、动力机制等角度解构了“代表”与“选举”之间的天然联系，反映出现实政治世界中代表关系的复杂样态。这些发展突破了政治代表的传统理论定式，勾勒出代表理论的多元图景，为我们思考算法世界的政治代表关系提供了理论支点。

本文接续“非选举型政治代表”的前沿思潮，尝试提出“算法代表性”的理想型政治概念，用于描述算法统治下不依赖正式选举机制的新型政治代表关系。简言之，算法代表性是指在算法系统的输入、运算和输出关键环节嵌入授权、问责和回应的实质性政治代表活动，以保障公民在算法统治下的基本权益和出场机会，限制“算法利维坦”的权力边界和运行方式，形成新技术环境下的新型社会契约和代表关系模式。

首先，算法代表性是一种理想型的规范概念，作为分析框架用于评估现实世界算法过程（输入、运算和输出）各个环节的代表性质量，规范算法场域新型代表关系的建构和完善，其作为目标指南是现实运作难以达成却可不断接近的理想状态。其次，算法代表性既注重实质性代表（substantive representation），又强调描述性代表（descriptive representation）。前者旨在算法过程中嵌入授权、问责和回应等实质性代表活动，后者关注脆弱群体、少数族裔、低收入阶层、老年人、女性等在算法过程中实现平等代表。再次，算法代表性归属于“非选举型代表”，它虽然无法依赖由选举产生的正式委托代理关系，但依然可以遵循前沿代表理论提出的偏好建构路径、话语代表路径、内驱代表路径、公民代表路径等，推动算法代表关系建构和代表性达成。正如雷菲尔德所指出的，代表关系的本质是被代表者对代表者的某种“接纳”，这种接纳可以通过选举的方式表达，也可以通过非选举的途径实现。[16]不能因为非选举竞争场景，而忽略政治代表问题。最后，算法代表性的提出一方面是运用政治学前沿理论解析和评估全新的算法统治现象；另一方面也借助算法代表关系的反思和建构，丰富新技术环境下的代表理论。

算法统治的代表性问题及其生成机理

智能算法深度嵌入到政府管理和社会治理当中，服务于识别、监管、预测、预警、协同、优化、决策、指挥等政务功能。政务算法化有利于减轻政府运行成本、改进政府决策质量、提升政府治理效能。然而，伴随算法沿着政务树不断爬升和扩散，进入更加高阶和广阔的公共决策领域，其开始动摇权利与权力、自由与秩序、公平与效率之间的平衡关系，引发算法统治的代表性危机。本节将研究视角从“运用算法的治理”转向“针对算法的治理”，借助算法代表性的概念和分析框架，考察政务算法系统的输入、运算和输出过程在授权、问责、回应维度存在的代表性问题及其生成机理（见图 1）。

（一）算法输入环节的代表性问题

算法统治始于数据，数据决定了算法系统的性能上限。为实现特定政府管理和社会治理目标，算法首先需要“数据喂养”，即数据输入。根据目标设定，政府及算法服务商开展数据收集和数据准备，包括用于训练算法模型的数据、目标函数的数据以及算法施用对象的数据。然而，无论是数据收集还是数据准备环节，当前都存在授权、问责和回应上的严重缺失，造成代表性错配或代表性偏差。

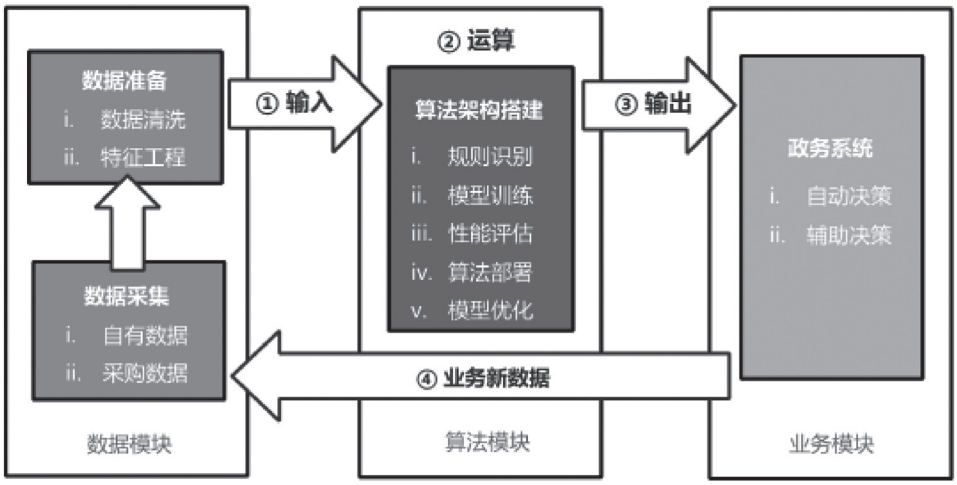

图 1　政务算法系统的过程示意图

在数据收集方面，政务算法聚合的大类数据源通常包括各类行政数据、社会运行数据、社会行为数据、人口学数据、社交媒体数据、监控图像数据，乃至物联网和身联网数据等，并根据运用场景和任务目标，抽取和匹配足量的相关数据。同时，还需满足任务目标函数的数据点，即用于识别模式和创建规则的可“标记”为正（例如社保舞弊）或负（例如社保非舞弊）的结果变量。公共部门的算法开发主要依托官方自有数据，即政府在日常业务开展和社会管理过程中积累的各类数据，例如社会福利数据、执法记录数据、城市运行数据、行政服务数据等，这些数据通过跨部门协同实现内部聚合。此外，向数据服务商（data broker）采购数据也日渐盛行。例如，美国的 SocialNet 公司为公共安全部门服务，宣称能够从全球 120 多个社交平台以及数据转储和 RSS 馈送中实时收集数据。当前西方国家已经形成规模庞大的数据交易市场，著名的政务数据服务商包括 Experian、i360、Aristotle、L2 等。数据收集是算法开发的前提，贯穿于政务算法运用和优化的全过程，算法系统部署后生成的新业务数据也将返回数据仓库，形成“永续数据采集”（permanent data collection）的循环。

在数据准备方面，为确保算法输入端的数据质量，须对原始数据进行一系列准备操作，特别是数据清洗（data cleaning）和特征工程（feature engineering）。首先，原生数据通常存在格式不统一、程序语言多样、重复性、非均衡性、非法值、缺失值、离群值、错误记录和特征依赖等质量问题，需要对其进行过滤、去重、插补、变换（标准化、归一化、缩放）、重置（集成、分割）、采样（分层、平衡）等数据预处理，将脏数据变成清数据，将非结构化数据转换为结构化数据。在此基础上，用于训练算法的数据需要组织成特征并实现降维降噪，以求更好地表达和描述数据，该过程即为特征工程。该环节需要政务专家团队、一线业务人员与算法工程师等结合领域知识（domain knowledge）进行数据探索，开展特征选择、提取、构造、测试和改进，创建用于特定算法目标的数据集（fit-for-purpose data）。数据准备通常占去整个分析管道的大半时间，虽可借助“自动数据准备”（ADP）框架和工具，但确保算法性能离不开专业领域的数据洞察和业务知识。

然而，算法的数据收集和数据准备阶段都面临代表性困境，体现在授权、问责和回应的全过程。首先，在授权方面，缺少数据权益方的授权程序，公民甚至并不知晓自己的哪些数据在什么

时候被哪些机构通过什么方式所记录、采集和流转。虽然当前普遍采用了“知情–同意”程序和“最少必要原则”，但是用户并没有太多实际选择空间和约束权限。在公共领域，数据更是被当作政府资源，数据收集被认为是默许的和正当的。诚然，当前有关数据确权问题尚存在较大争议，围绕所有权、使用权、变更权、收益权等数据权益划分缺少明确的法律认定，甚至在不同法制体系和应用场景中也存在差异。然而，从政治代表的角度来看，数据作为个人身份属性和要素资产的延伸，在其被用于特定政务用途时，实际产生了权利的交托和转移，该过程理应有实质性的授权程序。此外，虽然政府使用数据是服务于公共目的，但是算法将对公民个体的境遇和权益带来差别化影响，尤其在数据不完备时上马算法系统，授权程序是对算法统治风险的前置预防。

其次，在问责方面，数据权益方难以实施监督和问责。数据权益的让渡和转移随之产生数据责任。然而，算法输入阶段的数据采集和准备具有较强的隐匿性、专业性和渗透性，公民通常并不清晰知晓其数据被采集和使用的目的、方式和范畴以及数据质量问题带来的算法后果，因此较难对该过程的合法性、必要性和准确性展开监督和问责。一方面，责任主体不清晰，数据的采集者、交易者、处理者、使用者和管理者身份交杂，各主体间的法律责任不明晰，数据流通链条和交互过程复杂，使得权益人难以追溯和辨识相关责任方，加之各种合约自赋的免责条款，权益人较难施加控制和问责；另一方面，公民的权利主体地位模糊，较难依托法规主张其知情选择权、数据更正权、数据删除权、自动化处理反对权等。针对数据收集是否具有合法权威、数据采购是否来自合规渠道、原生数据是否完成脱敏隔离、数据是否反映业务现实等问题，处于权利弱势地位的公民个体难以将自己的偏好、利益和诉求传导进算法输入过程，无法行使有效的监督与问责。

再次，在回应方面，数据权益方难以获得救济和回应。数据收集和数据准备过程可能侵害数据主体的合法权益，例如，算法服务商超过合规性、正当性和最少必要原则采集数据，或对数据的存储、传输和管理不当，造成数据权益纠纷或安全风险。在数据准备时由于数据的代表性、测量的准确性、特征提取偏差等问题导致算法输入端的盲目性和歧视性，对目标群体产生权益影响。理论上，数据权益人或其监护人有权通过各类法律和业务渠道进行申诉。然而，实操中公民很难提出反对和寻求回应。一方面，数据收集、交易和处理的链条复杂，侵权主体相互避责推诿（特别是涉及算法外包服务时）；另一方面，数据权益人举证难度较大，算法行为与结果的因果关系难以证成；再者，政府和企业作为算法控制方，处在数据生态链中的优势地位，可以选择性回应或干脆不回应，并借由政务算法的公共属性排斥公民异议。

以欧美国家的预测性警务算法为例，为了推动警务执法从“应对型”（reactive）向“预测型”（predictive）转型，欧美国家正致力于推动“警务工作算法化”。[17]例如，荷兰阿姆斯特丹开发了CAS 犯罪预测系统，基于历史犯罪数据和实时动态数据预测城市犯罪风险，将警务资源和执法重点部署在预测风险较高的地区。德国柏林运用警务预测系统 KrimPro 将首都划分为约 4500 个治安网格，通过收集和挖掘各类数据预测治安风险，优化警力调配和安全防控。[18]美国洛杉矶警局成立了“实时犯罪分析中心”（Real-Time Crime Analysis Center），部署基于智能算法的犯罪预测系统，其数据采集和识别可以精准到人。超过 60 个美国城市在警务执法中接入了类似的算法工具（例如 PredPol 系统和 HunchLab 系统），这些系统聚合和导入了各类公民数据（包括视频监控数据、行踪数据、犯罪档案数据、消费和财务数据等）。虽然预测性警务算法系统有助于提升治安管理

和犯罪预防的效能，降低执法活动的公共预算，但是，其在数据收集和数据准备环节中存在严重的代表性问题。

警务算法在授权、问责和回应三个代表性维度存在缺失，数据收集并未经过数据权益人的授权，公民群体难以对警务算法开发者和使用者实施有效问责，权益遭受不公正对待的公民较难通过诉讼和申诉寻求回应。首先，预测性警务算法漠视公民的“知情–同意权”，强化了对公民个人的全景式监控。在无授权的情形下，公民的线上和线下活动数据无时无刻不被记录和观测，人类活动被置于机器的自动化算法统筹之下。[19]其次，数据质量低下或存在前置偏见，依赖“脏数据”进行特征提取，预测性警务系统会产生歧视性结果，加剧执法过程的不公正风险。研究发现，美国警务数据普遍存在记录错误、选择性录入、覆盖偏差、数据缺失、人为数据操弄等质量问题，[20]严重影响其分析的内部有效性和预测的外部有效性。这些不准确、扭曲和充满偏见的“脏数据”造成警务算法的模型偏差和预测失准，产生算法强化的公民权利侵犯。[21]此外，基于历史数据的算法系统存在身份、族群和外貌特征等方面的识别偏见，对某些特征字段极其敏感，其结果是“警务算法化”不仅没有让执法工作变得客观中立，反而加剧了对特定人群的算法压制。[22]再次，公民通常缺少对数据处理过程的认知能力，难以辨识和追溯具体责任人，执法机构倾向于将错误归咎于没有人格属性的算法，从而逃避问责。即便专业人士提出质疑，官方也置若罔闻。例如，2020 年 1400 名数学家联署，通过美国数学学会呼吁停止使用预测性警务算法和相关数据活动，但此举未能阻止更多的美国警局采用算法工具。最后，第三方算法供应商出于成本考虑也会拒绝承认和纠正数据收集和准备过程中的问题，导致权益人诉求无法得到回应和救济。

（二）算法中间环节的代表性问题

算法中间环节是根据所定义的中心任务对数据集进行训练，依托特定技术框架，在不同程度的人为干预或者完全摆脱人为干预的情形下，完成算法规则的设计、搭建、评估和优化，形成自我感知、学习、创造和行动的自动决策系统（automated decision-making system），最后将算法系统部署在政务应用场景当中。针对某项政务目标，通常存在多种潜在算法方案，算法系统搭建（解决方案的原型化和选型）的过程是不断逼近数据潜力和模型性能的过程。然而，该过程同样存在授权、问责和回应维度上的代表性缺失。

政务算法系统通常挖掘历史业务数据中的潜在关联模式，通过提取特征变量和构建复杂模型来识别和拟合训练集数据中的变量关系，并通过测试集和验证集数据优化模型状态，然后依据一系列规范标准（准确率、精确性、ROC 曲线等）进行系统性能评估，最后应用于样本外泛化和新对象的预测或分类。在部署和执行之前，各种算法方案需要进行性能评估和选型测试，通常涉及大量的实验（包括选取不同的特征变量、调整各个参数、更换算法框架），以验证其是否满足任务预期及其性能优劣。算法方案在应用过程中不断优化迭代，通常以接入新业务数据、增加数据覆盖率、提升数据集质量、调整变量和参数等方式实现。在政务领域，常用的算法底层技术包括机器学习、深度神经网络、自然语言处理、增强智能、图像处理、决策树程序、社会网络分析、模拟等。随着算法“物种大爆发”，系统的自我训练和学习能力不断提升，甚至可以摆脱人为设计和干预，完成从数据探索、算法构建、性能验证到任务执行的高度自动化。该进阶趋势一方面是对政务算法的强大赋能，另一方面复杂算法的技术原理越来越超出使用者、适用对象甚至开发者的理解和控制范畴，潜藏了算法代表性危机。

政务算法系统的自动化和复杂化产生普通民众甚至专业人士难以洞悉的“隐层”，造成权益人较难控制算法的搭建和运行过程，无法保障自身权益的表达、再现和落实，导致严重的不在场性，以及特定群体的代表性缺失和代表性错配。算法的模型建构和应用执行过程缺少授权、问责和回应的代表性保障，引发算法场域的决策专断和隐形霸权。

首先，在授权方面，政务系统是否适合引入算法框架，哪些数据用于算法训练、测试和验证，通常均未获得权益方的授权。算法设计依赖历史数据集的信息提取，即便权益人已被告知并同意数据收集行为，但并不等同于其授权数据的使用目的和范畴。虽然模型建构可以使用经过脱敏和加密处理的衍生数据（例如联邦学习和隐私计算），但是从原始数据到衍生数据的转换以及衍生数据的算法运用也需正当程序的确认。同时，历史数据集存在某些类型和特征的数据被过度记录而其他字段缺少记录的情形，引起模式识别、特征提取和参数赋值的偏差甚至错误，导致算法系统无法反映业务现实（ground truth）。在算法执行环节，公民通常也未授权算法调取自己的信息来自动生成影响自己权益和福利的决策结果，算法应用对象甚至不知道其面对的是算法自动决策机制而非人工服务。

其次，在问责方面，算法过程变得高度“黑箱化”，模型运行原理和预测过程的可解释性较低，导致问责困难。算法系统作出自动化决策并将结果指令传递到现实世界中，但对于算法系统的设计意图、技术逻辑、决策机制和责任归属鲜有公开。诚然，“算法黑箱”有一定的技术和伦理合理性。例如，黑箱是对复杂技术系统的简约化封装，以减轻使用者的采纳负担，形成简洁稳定的界面层；同时，“算法封装”对数字中的个人隐私、知识产权和系统安全提供了保护屏障。[23]但是，“算法黑箱”的负面效果日渐凸显，“算法封装”会造成严重的“规则隔音效应”，即公民 / 管理对象与决策者、算法设计者之间存在对算法规则的信息落差，使得合规审查和问责机制难以奏效。算法过程问责可分为事前问责和事后问责，内容涉及为何使用算法以及如何使用算法。事前问责旨在寻求对算法系统预期用途、设计逻辑、问题定义和目标设定的解释，以评估算法引入的必要性和合理性；事后问责关注算法应用过程中的工作原理、运行模块、执行性能、如何生成结果以及存在哪些应用限制。算法问责既寻求系统整体功能和决策程序的“全局性解释”（global interpretability），又寻求算法在具体场景下作用于单个样本或类别样本的“局部性解释”（local interpretability）。[24]对于公共部门，因为算法运用涉及公共资源的权威性分配，算法的问责标准应高于普通商业领域，其相关决策过程需要提供更加透明的可解释性框架。然而，现实中无论是算法的模型理论、学习路径、参数维度、变量交互、决策程序和控制范围等运行原理，还是特定场景的信息来源、决策规则、参考要素、权重占比以及决策生成路径等过程要素，都具有很强的隐匿性和复杂性，算法权益方要么无法触达，要么难以理解这些算法中间过程。特别是当深度神经网络等无监督、高度自主和动态变化的算法被引入时，其算法运行过程甚至会超出专业开发者的理解能力。此外，公共部门和算法服务商也倾向于以公共安全、商业机密、知识产权为由拒绝公开算法信息，导致“算法黑箱”风险难监管、难溯因和难问责。

再次，在回应方面，可分为针对算法全局系统适用性、合理性和合规性的回应，以及对局部场景和具体案例决策过程的回应。在全局回应方面，算法过程的黑箱化受到主观因素和客观因素的影响。就主观因素而言，算法采购者和设计者会将自己的价值理念、假设判断乃至利益企图植入到复杂的算法规则当中，并通过“算法封装”予以遮隐。在因目标偏移、设计不当和利益侵害遭受指控时，则归咎于非人格化的算法，以逃避回应。就客观因素而言，算法技术的

复杂化会超出开发者和使用者的理解范畴，加之问题定义、算法设计、模型训练、系统部署、采购交付等各个环节的分割，导致单个责任主体无法确知算法全过程的内部工作原理和决策生成程序，在算法产出非意图的后果时无法作出确切回应。在局部回应方面，因为算法的训练和建构是基于数据集层面的特征提取、模型拟合和性能验证，一方面其精确性和准确性只能保证在一定统计水平之上，无法实现所有个体层面的完全准确，另一方面算法模型存在对历史数据“过拟合”的风险，在模型外推而应用于新样本时容易导致“样本外”预测偏差。上述偏误属于学术研究可容忍的范畴，但是当算法部署于现实世界的公共决策场景时，这些偏误将落在具体个案和业务对象身上，导致个体层面的不公正和权益受损。更有甚者，算法系统并不负责确认个体数据的准确性，如果某个关键数据记录存在错误，算法则可能产生对该用户不利的结果。公民个体在寻求救济时，会面临算法素养欠缺和专业知识短板导致的能力差距，无法获得算法系统针对具体个案的有效回应，也无法行使反对权，要求放弃算法决策而改用人工服务。此外，算法赋予了处于技术和信息优势地位的公共部门以强大权力，面对算法霸权个体往往难以获得及时而充分的回应和救济。

以司法裁决算法系统为例，一些欧美国家在司法领域引入算法评估嫌疑人的逃离风险、累积犯罪风险和公共安全风险，以生成假释裁定甚至审判量刑。[25]此类系统通过挖掘历史犯罪记录、审判文书、刑事处罚数据中的要素关联模式，并结合人口学数据和社交媒体数据，提取特征变量和训练交互程序，预测再犯或逃逸的风险值（risk score）。该风险值（例如取值 0~10 分之间）可以辅助法官作出司法裁决，甚至自动作出是否假释或缓刑的智能裁判。算法系统有助于限制法官的主观裁量，维持统一的司法标准，缓解案件堆积压力，目前在美国已用于 29 个司法辖区，包括 3 个州全境（亚利桑那、肯塔基和新泽西）。[26]司法算法并不限于审前（pre-trial）的逮捕或释放决定，甚至广泛涉足审判活动。盛行的审判算法应用系统包括 LSI-R 和 COMPAS，两者均由商业公司开发，前者应用于加利福尼亚和华盛顿等州，后者被密歇根、佛罗里达和威斯康星等州采购使用。其中，COMPAS 系统基于 5 个关键维度的变量数据进行建模，包括犯罪活动、社会关系和生活方式、性格和态度、家庭状况、社会排斥，通过综合静态和动态因子建构风险预测算法，辅助和指导具体的量刑裁决。[27]

就算法中间过程而言，司法裁决算法在授权、问责和回应三个代表性维度存在显著缺失。首先在授权上，刑事裁判算法在训练、测试和运行过程中的数据运算和模型选择行为并未获得公民授权。算法供应商根据服务协议远程调取多个公共部门的数据进行模型训练，而这些服务协议缺少授权程序。为何引入算法架构，哪些部门接入了算法系统，哪些数据用于模型搭建，以及谁决定部署应用等都未经过授权，而被化约为司法部门采购算法服务的官僚程序。在问责上，司法裁决算法宣称融合了犯罪学、社会学、心理学理论与数据科学技术，但其在历史数据的模式识别、预测变量的特征提取、算法模型的学习路径、核心变量的交互关系、风险值测算中的权重分布等方面可解释性较差。当算法系统当作服务产品被交付和部署后，法官、律师、嫌疑人和其他相关方都无法追踪其工作原理和决策过程。例如，决定假释或监禁界限的阈值（处在 0~10 分之间的某个风险门槛值）如何划定？不同阈值下算法预测的准确性如何？针对具体行为人的处罚决定如何作出？相关技术过程通常秘而不宣，给全局性和局部性的问责带来挑战。在回应上，司法裁决算法是基于历史数据集训练的总体预测模型，其将人群按照不同属性和特征组合进行聚类，缺少个体化的案件评判和裁量，样本外偏差也容易导致个案层面的报错和误判；算法裁决依据的是历

史上同类人群的犯罪倾向，而非现实中具体个人的实际犯罪行为；当训练集数据存在错误和偏见时，将对某类群体（例如少数族裔）带来系统性不利影响或产生个案层面的非意图后果。[28]算法将司法过程模式化和标准化，个体无法参与、矫正、阻止或寻求算法救济，算法的强技术属性导致司法部门难以纠偏和对个体申诉作出回应。

（三）算法输出环节的代表性问题

基于算法的公共决策依赖算法的最终输出，输出的结果形式、准确性、稳健性、公平性以及输出的应用转化机制将对政务领域的算法决策效能产生决定性影响。无论算法中间过程多么复杂，其必须在执行有限步骤后进入终止状态，生成确定性的输出结果，该数字结果根据算法系统的任务设定转化为自动决策的指令或辅助决策的参考。在政务场景中，算法输出环节至少存在三个关键要素。

其一，输出的目标选定，即定义算法系统的中心任务，是为了提升效率还是实现公平，同时还需权衡投入成本、安全性能、系统复杂度等。算法具有任务单一性特征，无法同时达成上述全部目标，不同目标之间本身也存在冲突，为确保中心任务可控必须作出取舍，而这种取舍会涉及公民的社会经济权利。

其二，输出的可解释性，即算法系统能否为其输出结果提供有意义且可理解的解释说明。在政务场景中，算法输出的不仅是字符串，更是有决策意义的结果信息。算法结果有时甚至是反直觉的（counter-intuitive）和超常识的，需要以可理解和可说服的语言传递给外界。例如，输出的主要影响要素有哪些，某个特定输出的决定要素是什么，改变该要素是否会产生不同的输出，为什么不同案例会产生相同的输出，为什么相似案例会产生不同的输出，为什么同一案例在不同时间会产生不同输出，这些问题都关乎输出结果的合法性以及公众对算法系统的信任接纳程度。[29]

其三，输出的场景化应用，即从算法输出到公共决策的机制设计，将算法系统部署到业务系统当中实现产品化。算法架构与业务架构的匹配度如何，人工干预与算法输出如何协作，业务系统对算法失准的容忍度如何，怎样处理算法标准化与场景复杂性之间的矛盾，算法架构能否满足未来业务发展和转型的需要，这些问题涉及算法输出以何种形式影响公共决策和产生怎样的社会后果。由此可见，算法输出关涉价值设定、业务交互以及社会影响评估。政务算法必须纳入算法输出的业务意义、社会意义和政治意义的考量。

算法输出是算法系统发挥治理效能和影响公共决策的关键环节，其与具体政务场景结合将对公民的基本社会经济权利产生直接和深远的影响。然而，当前公民的利益和声音在该环节难以得到表达和再现，存在严重的不在场性，缺乏实质的代表性保障。

首先，在授权方面，算法输出在目标设定和应用转化上缺少事前授权。在目标设定上，算法通常专注于效率，而忽略体验、安全和公平；即便追求公平，不同算法系统对公平的技术定义和实现路径也存在差异。[30]算法目标通常根据公共部门需求设定，追求业务运行的智能化和决策过程的自动化，从而节省业务成本并提升治理效能，其他目标（例如公平、准确、隐私、安全）可能被边缘化。如果优先考虑公平，那么如何对公平进行定义和操作化将决定算法目标函数的优化方向，其间的价值和权益取舍鲜有授权。在应用转化上，算法输出是接入业务端口直接转化为自动化决策还是由人工干预辅助决策，谁来决定算法结果以何种方式影响公共决策，这些过程都缺少算法适用对象或公众利益代表的参与和协商，产生单方面抉择的代表性困境。

其次，在问责方面，针对算法输出的社会影响缺少问责机制。由于数据和模型偏差，算法输

出存在忽略和压制特定人群、瞄准和威胁特定人群、筛选特定利益和人群、放大特定偏好和政策的风险，产生对某些族群、性别、阶层、信仰群体的区别对待甚至结构性歧视。[31]算法在非预见性和样本外的情境下也会产生涌现性的偏见（emergent bias）。同时，算法开发者对任务目标的前置理解和技术表达可能并不符合政府的业务初衷，也将带来非意图的社会损害或者全局性负面影响，导致“算法失灵”。此外，政务领域也无法排除对算法输出结果的恶意利用。然而，目前社会缺少对算法权力的制约和监督，难以获得算法输出的解释性说明，也无法倒逼算法开发者和使用者采取有效措施防范相关风险、歧视和偏差。在算法与人工交互的决策场景中，算法输出与人工干预的责任划分不明确，导致溯因和追责困境。当业务场景高度依赖非透明的算法决策时，公共部门倾向于将责任推诿给上游的算法供应商甚至是不具法人资格的算法系统，算法反而成为新的“避责”工具。

再次，在回应方面，算法输出的非准确性会对群体和个体的权利构成侵害。算法系统的诸多特性决定了其输出结果不可能达到完全准确，成本预算的约束、数据的量和质、特征的提取和构造、模型的训练和选择、样本外误差、目标函数的定义、决策阈值的选择等都是影响算法输出准确性的关键因素。[32]单就工程技术而言，算法系统可以在数据量和质不足的情况下提前启动（所谓“数据不够，模型来凑”），通过浅层模型、快速学习、强化学习等技术手段最大化利用“脏数据”，并借助新业务数据不断进行修复和优化。然而，落实在具体政务场景中，这意味着算法输出的准确性在较长时间内难以得到系统提升，导致经常性误判、误伤、报错甚至无法运行。加之业务领域的季节性因素和突然状况，带来个体和群体权益侵害的风险加大。算法输出的错误风险和个体权益的算法敏感性决定了必须提供制度化的回应机制，从而保障权益人在遭受算法不确定性时获得替代方案和救济回应。此外，算法应用还须考虑社会情境，算法输出会带来经济社会权益的再分配，甚至改变社会中既得利益格局，容易引发社群冲突，这时弱势方需要得到官方的救济和回应，以确保算法目标达成。当前，由于算法决策者、采购者、设计者、使用者的角色隔离和权责模糊，导致了算法输出环节的回应和救济困难。

以社会福利监测算法为例，欧美国家近年积极推动“数字福利国家”（digital welfare states）建设，大量运用算法技术监管社会福利系统，尤其是识别和打击福利欺诈。荷兰是老牌福利国家，为了缓解福利支出的负担和提升福利供给的精准度，政府引入了基于智能算法的风险探测系统（risk indication system，SyRI），用于监测和惩罚福利项目中的失范行为。该系统可以跨部门调用公民就业、税收、居住、教育、资产、债务、失业、移民、行政处罚等数据，并匹配公民的人口学数据、精神健康数据、社会活动数据等，通过算法模型预测福利申请人的作假嫌疑，输出公民的作弊风险指数（fraud risk score）。[33]任何有意打击福利欺诈的政府部门（例如社会保障部门、劳动就业部门、市政当局）都可使用这套系统，并根据其输出的监测警报和嫌疑信息（flagged citizens）启动精准调查和执法。该算法监测活动通常以整个社区为单位开展，即被监测社区内的所有公民（无论是否参与某项福利分配）都要接受该系统的检核和评估，该过程及其结果都无需告知民众。算法系统在数据分析和模型预测环节会对敏感信息进行脱敏处理，算法输出的结果是高风险福利舞弊嫌疑人的加密身份码列表，该列表可以对应真实的姓名和身份，从而方便相关部门开展瞄准式的执法活动。

福利监测算法虽然能够帮助政府减少财政浪费和提升分配效能，但是其算法输出环节存在授权、问责和回应三个代表性维度上的显著缺失。首先，在授权上，算法输出目标是为了识别

福利舞弊，其预设某些特征的社区和人群存在较高舞弊风险，而输出结果直接会触发福利部门的调查和执法活动，对这些社区和公民的福祉带来重大影响。在针对某些社区启用 SyRI 风险探测系统时，政府试图诉诸法律依据以规避社区居民的授权，[34]导致差别对待和算法误用。例如，SyRI 算法系统被发现主要针对低收入社区和双重国籍人士，在没有任何授权的情况下将部分人群置于更加严苛的算法监视之下。数字福利国家因此被批评为没有公共授权的压制性福利国家（repressive welfare state）。在问责上，SyRI 算法系统在没有监督和问责的环境下秘密运行。2017 年荷兰政府曾拒绝公开 SyRI 系统的算法模型，由此算法输出的生成过程和依据要素沦为“黑箱”，产生对特定群体（移民家庭、少数族裔和底层民众）的偏见、歧视和压制。此外，算法系统的开发者、运维者、使用者是分开的，调用算法的政府部门并不熟知算法的运行原理，最后依据算法结果开展执法的又是另一批人，当执法工作引起社会不满时，相关部门相互推诿甚至将责任推卸给非人格化的算法工具，加剧了问责难度。在回应上，SyRI 算法系统的输出结果并非完全准确，福利部门据此开展事后惩罚，存在严重的执法不公，对此缺少救济和回应渠道。例如，申请福利补贴的家庭由于不善于文书和电子申报，导致申请材料填写和签名等方面的行政错误，却被算法识别为欺诈，导致高额罚金甚至刑事定罪。此外，模型赋予某些特征过高的决策权重导致识别偏误，例如算法根据家庭用水量数据来判定单身人士的住房补贴舞弊。相关受害者申请救济非常困难，一些家庭因为不公的行政处罚甚至导致家庭破产和社区歧视。[35]总之，由于缺少授权、问责和回应的代表程序设计，算法系统在应用于福利领域时通常会造成公民权益侵害。[36]

结论

算法统治是正在形成中的新型国家治理形态，政府借助不断精进的算法技术管理公共事务和开展公共决策。近年来，基于算法的智能管理和自动化决策无论在规模上还是在密度上都在迅猛拓展，[37]不断增强国家的规制性权力、强制性权力和分配性权力，重塑着国家与社会关系以及政民互动形态，形成数字时代新的政治代表场域。诚然，算法对政府管理和社会治理具有强大的赋能效应，推动了政府体制内部的管理模式转型和业务流程再造，同时极大提升了社会治理和公共服务的效率效能。然而，政务算法化势必影响政府的权力构成、公民的权利实现以及政民关系模式，有必要从政治学理论的高度重新审视和研判算法时代的政治生活。本研究从政治代表性的前沿理论视角出发，提出“算法代表性”的概念和分析框架，审视和解析当前算法统治的代表性困境，探索非选举的治理场景中政治代表性的实现机制。

本研究发现，政务算法化正在推动新型统治形态和治理模式的形成，当前算法统治距离理想型的算法代表性尚存在较大差距。首先，在算法过程的输入、运算和输出环节都存在授权、问责、回应方面的实质性代表性问题（见表 1），造成公民声音的不在场性和公民权益的损失风险。其间虽有算法的技术特性使然，然则更重要的是缺少算法过程政治代表性的体制机制设计和操作行动实践。虽然政务算法场景的授权、问责和回应难以通过竞争性选举的委托代理机制直接达成，但这并不意味着实质性代表在算法过程中可以被忽略。相反，作为典型非代议非选举的治理场景，诸多非选举型的代表机制（例如偏好建构、话语代表、内驱代表、公民代表）依然可以被创制和实践，促成授权、问责和回应等代表性目标的实现。因此，面对“算法利维坦”，需要从算法代表

性的视角，重新设计和建构算法过程的公民出场和权益再现机制。

其次，在算法过程的输入、运算和输出环节同样存在严重的描述性代表问题，影响了脆弱人群、少数族裔、低收入阶层、老年人、女性等在群体和个体层面的权益平等实现。在算法统治之下，现实世界某些社会群体面对的歧视、偏见、压制和不平等在治理过程中被数字记录和储存，形成数据层面的歧视、偏见、压制和不平等；数据被调取、学习和训练，输入算法模型进而固化为算法层面的歧视、偏见、压制和不平等；算法输出及其业务运用产生政务决策，外化为决策层面的歧视、偏见、压制和不平等；政务决策带来治理后果和社会影响，衍生和累积形成现实世界更深的歧视、偏见、压制和不平等（见图 2）。如此循环往复，纯粹技术驱动的算法统治会不断沉积、固化、强化乃至结构化特定群体的弱势地位，排斥和压制少数群体的平等出场和权益实现。因此，面对政务算法化，需要从算法代表性的视角创新非选举型代表机制，推动弱势群体的权益保障。

表 1　算法统治的实质性代表性问题

	授权	问责	回应
输入环节	数据收集和准备过程中的数据权利让渡和转移缺少授权程序	数据责任体系不完备；数据责任主体不清晰；数据权益主体地位模糊；问责机制缺失	数据权益救济过程复杂；回应性面临算法权力的排斥
中间环节	对算法系统如何处理数据、如何训练模型、如何生成预测、如何将公共事务代码化缺少授权程序	算法黑箱化；全局可解释性和局部可解释性差；事前问责和事后问责困难	针对全局系统适用性、合理性和合规性的回应力差；针对局部场景和具体案例决策过程的回应力差
输出环节	算法输出在目标设定和应用转化上缺少事前授权	对算法输出的区别对待和结构性歧视、算法失灵和非意图后果、算法被恶意利用等缺少问责机制	面对算法输出错误、算法结果的不确定性、算法造成的社会利益冲突，缺少救济和回应机制

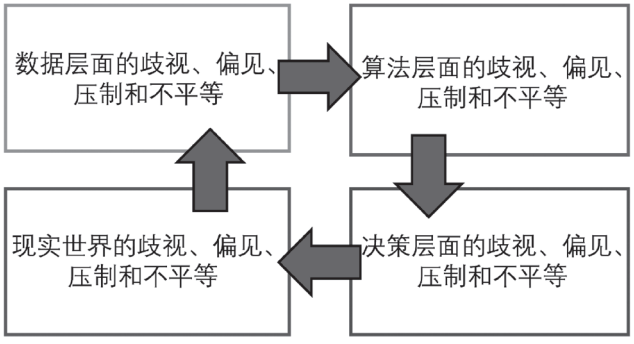

图 2　算法统治的描述性代表性问题

政务算法涉及公共权力的行使和公共资源的分配，其代表性标准应高于其他算法应用领域。当前，算法治理长期依赖自上而下的监管模式，而忽略自下而上的代表模式。在数字时代建构算法代表性，可以同时遵循内在进路和外在进路。就内在进路而言，可以在国家体系内部设置公民算法权利的代表机构或职位，例如议会中的算法审计和合宪性审查委员会、算法影响力评

估委员会，政府中的“首席算法官”。[38]就外在进路而言，可以推动关注算法治理问题的公民倡议团体、学术研究机构、第三方审计机构、公共媒体、算法听证、算法吹哨人、算法意见领袖、算法公益律师[39]等非选举型代表机制的建设。此外，在内在进路与外在进路之间搭建算法代表性的实践场域，例如算法的审议和授权平台、算法的监督和问责平台、算法的申诉和回应平台、算法的公共协商平台等，从而从授权、问责和回应维度对算法统治输入、运算、输出的全过程开展代表性约束。上述算法场景的代表性实践，也将极大丰富非选举型政治代表理论的解释路径和理论意涵。

随着数据的全量全要素连接和物理世界的数字化表达，数字驱动的算法统治将越来越普遍地统摄政务领域和公共生活。我们在推进政务算法实现治理功能的同时，还须重视其对国家与社会关系、政府与民众关系、权力与权利关系的深远影响。未来研究可以关注不同群体差异化的代表性诉求、不同算法应用场景的代表机制设计、不同代表机制的实践操作及其效能评估等。总之，算法代表性既是规范算法统治的理论范式，又是引导算法场景中代表机制建设的目标蓝图。以政治代表性的视角加强算法过程的授权、问责和回应制度建设，保障弱势群体的算法权益，有助于提升算法统治的合法性，增进民众的算法信任以及推动算法向善。

注释：

① Helen Margetts and Cosmina Dorobantu, “Rethink government with AI,” *Nature*, vol.568, no.7751, 2019, pp.163-165; Serge Abiteboul and Gilles Dowek, *The Age of Algorithms*, Cambridge: Cambridge University Press, 2020.

② John Danaher, “The Threat of Algocracy: Reality, Resistance and Accommodation,” *Philosophy & Technology*, vol.29, no.3, 2016, pp.245-268.

③ David Freeman Engstrom et al., “Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies,” *NYU School of Law, Public Law Research Paper*, no.20-54, 2020; Karen Levy, Kyla E. Chasalow, and Sarah Riley, “Algorithms and Decision-Making in the Public Sector,” *Annual Review of Law and Social Science*, vol.17, no.1, 2021, pp.309-334.

④ Kathleen Creel and Deborah Hellman, “The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems,” *Virginia Public Law and Legal Theory Research Paper*, no. 13, 2021; 张爱军：《“算法利维坦”的风险及其规制》，《探索与争鸣》2021 年第 1 期。

⑤ Dario Castiglione and Johannes Pollak, *Creating Political Presence: The New Politics of Democratic Representation*, Chicago: University of Chicago Press, 2019.

⑥ Hanna Pitkin, *The Concept of Representation*, Los Angeles: University of California Press, 1967.

⑦ Mark Warren and Dario Castiglione, “The Transformation of Democratic Representation,” *Democracy and Society*, vol.2, no.1, 2004, pp.5-22.

⑧ Lisa Disch, “The Constructivist Turn in Democratic Representation: A Normative Dead-End?” *Constellations*, vol.22, no.4, 2015, pp.487-499.

⑨ Michael Saward, *The Representative Claim*, Oxford: Oxford University Press, 2010.

⑩ John Dryzek and Simon Niemeyer, “Discursive Representation,” *American Political Science Review*, vol.102, no.4, 2008, pp.481-493.

⑪ Rousiley CM Maia, “Non-electoral Political Representation: Explaining Discursive Domains,” *Representation*, vol.48. no.4, 2012, pp.429-443.

⑫ Nadia Urbinati, “Representation as Advocacy: A Study of Democratic Deliberation,” *Political Theory*, vol.28, no.6, 2000, pp.258-786.

⑬ Jane Mansbridge, “A Selection Model of Representation,” *Journal of Political Philosophy*, vol.17, no.4, 2009, pp.369-398.

⑭ Laura Montanaro, “The Democratic Legitimacy of Self-appointed Representatives,” *The Journal of Politics*, vol.74, no.4, 2012, pp.1094-1107.

⑮ Mark Warren, “Citizen Representatives,” in Mark Warren and Hilary Pearse, eds., *Designing Deliberative Democracy: The British Columbia Citizens' Assembly*, Cambridge: Cambridge University Press, 2008, pp.50-69.

⑯ Andrew Rehfeld, “Toward a General Theory of Political Representation,” *The Journal of Politics*, vol.68, no.1, 2006, pp.1-21.

⑰ Andrew Guthrie Ferguson, *The Rise of Big Data Policing:*

Surveillance, Race, and the Future of Law Enforcement, New York: NYU Press, 2017.

⑱ Albert Meijer, Lukas Lorenz, and Martijn Wessels, "Algorithmization of Bureaucratic Organizations: Using a Practice Lens to Study How Context Shapes Predictive Policing Systems," *Public Administration Review*, vol.81, no.5, 2021, pp.837-846.

⑲ Sarah Brayne, "Big Data Surveillance: The Case of Policing," *American Sociological Review*, vol.82, no.5, 2017, pp.977-1008.

⑳ Andrew G. Ferguson, "Policing Predictive Policing," *Washington University Law Review*, vol.94, no.5, 2017, pp.1109-1189.

㉑ Rashida Richardson, Jason M. Schultz, and Kate Crawford, "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice," *NYU Law Review Online*, vol.94, no.15, 2019, pp.15-55.

㉒ Aleš Završnik, ed., *Big Data, Crime and Social Control.*, London: Routledge, Taylor & Francis Group, 2017.

㉓ Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," *Big Data & Society*, vol.3, no.1, 2016, https://doi.org/10.1177/2053951715622512.

㉔ Christoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, Lulu.com, 2020; Isabelle Guyon, et al., eds, *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Cham: Springer International Publishing, 2018.

㉕ Aleš Završnik, "Algorithmic justice: Algorithms and big data in criminal justice settings," *European Journal of Criminology*, vol.18, no.5, 2021, pp.623-642.

㉖ Danielle Kehl, Priscilla Guo, and Samuel Kessler, "Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing," Berkman Klein Center for Internet & Society, Harvard Law School, 2017, p.10.

㉗ Eugenie Jackson and Christina Mendoza, "Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not," *Harvard Data Science Review*, vol.2, no.1, 2020, https://doi.org/10.1162/99608f92.1b3dadaa.

㉘ Cynthia Rudin, Caroline Wang, Beau Coker, "The Age of Secrecy and Unfairness in Recidivism Prediction," *Harvard Data Science Review*, vol.2, no.1, 2018, https://doi.org/10.1162/99608f92.6ed64b30.

㉙ Min Kyung Lee, "Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management," *Big Data & Society*, vol.5, no.1, 2018, https://doi.org/10.1177/2053951718756684.

㉚ 参见 Shira Mitchell, et al., "Algorithmic Fairness: Choices, Assumptions, and Definitions," *Annual Review of Statistics and Its Application*, vol.8, 2021, pp.141-163; David Weinberger, "How Machine Learning Pushes Us to Define Fairness," *Harvard Business Review*, vol.6, 2019, https://hbr.org/2019/11/how-machine-learning-pushes-us-to-define-fairness。

㉛ Nima Kordzadeh and Maryam Ghasemaghaei, "Algorithmic Bias: Review, Synthesis, and Future Research Directions," *European Journal of Information Systems*, vol.31, no.3, 2022, pp.388-409; Jenna Burrell and Marion Fourcade, "The Society of Algorithms," *Annual Review of Sociology*, vol.47, 2021, pp.213-237.

㉜ Stefan Buijsman and Herman Veluwenkamp, "Spotting When Algorithms Are Wrong," *Minds & Machines*, 2022, https://doi.org/10.1007/s11023-022-09591-0.

㉝ 参见"算法监督"网站：https://algorithmwatch.org/en/syri-netherlands-algorithm，2022 年 4 月 15 日。

㉞ 参见荷兰相关立法：https://wetten.overheid.nl/BWBR0013267/2022-04-01，2022 年 4 月 15 日。

㉟ 参见荷兰媒体 Trouw 的系列调查报告：https://www.trouw.nl/tag/syri，2022 年 4 月 15 日。

㊱ Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*, New York: St. Martin's Press, 2018.

㊲ Zeynep Engin and Philip Treleaven, "Algorithmic Government: Automating Public Services and Supporting Civil Servants in Using Data Science Technologies," *The Computer Journal*, vol.62, no.3, 2019, pp. 448-460.

㊳ 例如，2019 年美国纽约市政府设立了"首席算法官"（Chief Algorithms Officer）的新职务，代表市民行使算法监督。

㊴ 例如，为弱势群体在算法福利分配中提供法律援助（Lawgorithms），参见 Michele E. Gilman, *Poverty Lawgorithms: A Poverty Lawyer's Guide to Fighting Automated Decision-Making Harms on Low-Income Communities*, Data & Society Research Institute, 2020, https://datasociety.net/library/poverty-lawgorithms/, May 25, 2022。

编辑　**杜运泉**