

算法歧视的认定标准

宁 园

摘 要 学界有关算法歧视规制的研究,欠缺算法歧视的认定这一逻辑前提,导致反算法歧视的规制范围过宽,算法应用与算法正义之间的矛盾被过分夸大,有必要明确算法歧视的认定规则。算法歧视以横向权力关系的存在为必要前提,不存在横向权力关系则算法歧视不成立。当算法控制者以用户固有的、难以改变的特性对其进行分类,并单方决定适用于用户的算法规则,用户只能被动接受决策结果而无退出算法系统之自由时,横向权力关系形成。在算法歧视的具体认定上,算法歧视以区分决策为行为要件,以用户遭受差别不利对待为结果要件。算法控制者若能证明差别对待具有合理性,则不构成算法歧视。搜索算法、人机对话算法对社会歧视的单纯呈现,以及社会经济发展、算法应用普及不平衡导致的算法红利分配失衡,均因欠缺区分决策要件,不构成算法歧视;“大数据杀熟”、评分算法所造成的差别对待并非必然构成歧视,须进行个案分析。

关键词 算法控制;算法歧视;横向权力;区分决策;差别不利后果;“大数据杀熟”;评分算法

中图分类号 D922 **文献标识码** A **文章编号** 1672-7320(2022)06-0154-12

基金项目 中国博士后科学基金特别资助项目(2021T140721);中国博士后科学基金面上资助项目(2020M680794)

随着算法应用的普及,算法歧视早已成为不可忽视的社会问题。学界围绕如何规制算法歧视,已经形成丰富的治理方案。然而,作为算法歧视规制的前提问题,“何种算法决策构成算法歧视”始终未被专门讨论,不少论者似乎都默认,凡是个人在算法应用中遭遇的不利对待均属于算法歧视。算法技术是现今社会发展的重要驱动力,严格规制算法歧视的同时,也不应任意扩张算法歧视的范围,过分苛责算法控制者。因此,为防止算法应用遭到过度声讨,本文尝试解决算法歧视的认定问题,以填补算法歧视规制的逻辑前提,促进算法规制方案的真正落地。

一、算法歧视认定环节的缺失现象

根据既有文献,学界对算法歧视的技术特点、产生原因、规制路径已有诸多讨论,并就以下四个方面基本达成共识:一是算法歧视现象频繁发生,已对社会平等正义造成巨大冲击,规制算法歧视具有现实必要性。二是算法控制者凭借算法技术取得了对用户的实际支配力,即算法权力;算法权力规制则是算法歧视规制的症结所在。三是算法并非完全中立,算法歧视既可能源于算法设计者的偏见植入,也可能源于已被现实偏见污染的分析数据。四是主要的算法歧视规制措施包括限制算法权力、构建算法透明机制、开展算法审查、赋予个人算法解释请求权等^[1](P39-40)]^[2](P1416)]。

遗憾的是,学界有关算法歧视规制的研究似乎预设了算法歧视认定的完成。在现有研究成果中,作为规制对象的算法歧视要么是一种抽象风险,要么是几种零星单薄的现象列举,而何种算法决策构成算

法歧视这一基础性问题遭到忽视。正是由于缺乏算法歧视的认定这一逻辑前提,学界有关算法歧视规制的研究总体停留在风险控制层面,难以向具体规制层面推进。尽管已有学者就算法歧视的具体民事责任展开论述,但其直接忽略了作为责任成立基础的算法歧视认定问题,实践意义十分有限^[3](P55-68)。

算法歧视认定环节的缺失,还进一步导致了算法歧视泛化的现象,突出表现在:一是将算法对社会歧视的单纯呈现泛化为算法歧视,如将算法搜索结果显示的歧视性语言,定性为算法控制者实施的算法歧视;二是将算法决策的一切差别对待泛化为算法歧视,如将合理的差别定价视为歧视性的“大数据杀熟”;三是将算法技术发展引发的社会不平等现象泛化为算法歧视,如将算法红利分配失衡导致老年人群体被边缘化定性为算法歧视。在算法歧视泛化趋势下,合理的算法决策以及非由算法决策导致的不平等现象,均被归咎于算法控制者^①(P29-30),这不仅会不当加重算法控制者的责任,更会激化算法利用和算法安全的价值对立,限制算法技术的发展和算法应用的普及。

当前,反算法歧视应从风险控制转向具体规制。为遏制已经出现的算法歧视泛化趋势,有必要明确算法歧视的认定规则。本文认为,算法歧视是算法控制者利用算法技术实施的、以算法决策为实现形式的歧视行为。算法歧视仍然秉承传统歧视的根本结构,认定时不应随意扩张算法歧视的范围;同时,算法歧视相比传统歧视亦具有不可忽视的隐蔽性和复杂性,在具体要件判定上也应作相应调适。循此基本理念,下文将着重围绕算法歧视的认定这一主题,阐明算法歧视发生所需的横向权力关系前提,算法歧视的行为要件和结果要件,以及算法歧视的构成例外。

二、算法歧视发生的必要前提:横向权力关系的形成

纠正算法歧视泛化,首先须明确反算法歧视以横向权力关系存在为必要。尽管学界普遍认可算法权力是算法歧视发生的重要原因,但并未将此提升至构成要件的地位。多数学者要么对算法歧视仅发生于横向权力关系中缺乏认识,要么倾向于将算法控制者与用户之间的一切关系预设为权力关系。既然认识到算法权力是算法歧视发生的原因,那么明确算法歧视发生以横向权力关系存在为必要,还需进一步阐释以下两个方面:一是算法歧视发生且仅发生于横向权力关系中;二是算法控制者与用户之间不当然存在权力关系,横向权力关系有其特定的形成条件。

(一) 歧视发生的必要前提:权力关系的存在

平等权作为一项基本权利,普遍为各国立法所确认。反歧视则是平等权的规范性表达,其禁止基于民族、种族、性别、宗教信仰、健康状况等特征对特定群体或个人进行不合理的差别不利对待^②,进而发挥限制公权力和保障社会弱势群体的作用。从传统实践来看,反歧视集中于公法和社会法领域,鲜见于私法领域^③。其原因在于,歧视发生以权力关系的存在为必要,而私法关系具有平等性,通常与权力无涉。

首先,从歧视的发生机制来看,权力关系是歧视形成的必要前提。依照马克斯·韦伯对权力的解释,当强势一方可以贯彻自己的意志而无需顾及弱势一方的反对时,即存在权力关系^[5](P264)。歧视是歧视主体在资源分配中对被歧视主体的利益限制和剥夺,其发生过程是歧视主体与被歧视主体之间的博弈过程,其实现以歧视主体可以强行贯彻自身意志为必要。在公法和社会法关系中,不乏掌握权力的强势分配者,反歧视则是防止分配者滥用分配权力、维护分配正义的重要制度工具。相反,在私法关系中,各主体之间依意思自治进行资源交换,歧视行为欠缺权力支撑,一方难以将其资源分配意志强加于另一

① 本文所称的“算法”是指人机交互决策的一系列步骤,即算法控制者通过代码设置、数据运算和机器自动化进行决策的过程,既包括人工智能算法,又包括传统计算机软件算法。算法控制者即通过算法进行决策的主体。

② 我国不少立法都有禁止歧视规则,具体可参见《就业促进法》第三条、《劳动保障部卫生部关于维护乙肝表面抗原携带者就业权利的意见》《广告法》第九条、《妇女权益保障法》第二条、《老年人权益保障法》第三条等。

③ 我国有关歧视的司法案例主要为就业歧视,具体表现为用人单位因劳动者的性别、地域、身高、患有乙肝或艾滋病等情形实施的就业歧视。除此之外,农村集体经济组织对女性、外来户、外嫁女在财产分配上的歧视在司法实践中也较为常见。

方。可见,由于欠缺歧视发生所必须的权力关系,私法关系通常与歧视无涉。

其次,反歧视规制也应以权力关系为限,在介入私法关系时保持谦抑,以避免侵蚀私法自治。理由在于,将“相同情况相同对待,不同情况不同对待”的诫命注入私法关系,无疑与私法自治背道而驰。在以意思自治为价值根基的私法领域,私主体间的关系由私主体自由安排,即使存在一定程度的偏见和差别对待,也多应归于当事人的行为自由范畴,如在“刘某诉中信银行股份有限公司信用卡纠纷案”(北京金融法院[2021]京74民终942号民事判决书)中,法院认为被告中信银行针对女性推出的信用卡申领专享服务,是在金融消费者需求多元化、差异化的基础上所设定合理的信用卡申领条件,并不构成歧视。况且,个人偏见源于生物本能、心理状态、社会文化等因素,具有不可根除性。个人行为普遍受偏见影响,对带有偏见的行为一概禁止是对行为自由的根本颠覆。

综上所述,反歧视规制以权力关系存在为前提,私法关系具有平等性,通常不会发生歧视,也很少受反歧视制度的约束。然而,算法技术的出现,尤其是算法区分决策的规模化应用,使算法控制者取得了事实上的资源分配权力,算法控制者和用户之间的平等关系也因此异化为横向权力关系。此时,歧视性分配的发生风险大大增加,算法歧视由此成为私法关系中亟待解决的新问题。

(二) 横向权力关系的形成要件:用户意志自由的缺失

算法歧视集中爆发于算法控制者与用户之间形成的横向权力关系,此一结论已得到学界广泛认同。然而,权力关系如何形成,则多见于技术层面的宏观描述,鲜见于要件层面的微观分析^[1](P34-36)。显然,算法控制者掌握算法技术并不等于其就取得或滥用了算法权力,权力关系是否形成,还应着重关注具体的技术应用是否符合特定要件。

权力关系与平等关系之间的本质区别在于,关系中的各方主体是否享有平等的意志自由。在权力关系中,分配仅体现分配者的意志,其他主体是受其意志支配的被分配者;在平等关系中,分配则是各方主体平等协商达成一致结果的过程。可以说,权力关系的形成,就是关系主体一方或多方意志自由被剥夺的过程。因此,在算法控制者与用户之间,横向权力的形成以算法控制者剥夺用户的意志自由为核心要素。

从博弈视角来看,平等双方在博弈中均享有意志自由,是否进行博弈(如是否订立合同)、博弈形成的规则(如具体的合同条款)以及由此产生的博弈结果不受任意一方操纵,即便一方最终承受客观的不利结果(如一方依合同以高于市场价的价格购得标的物),该方也始终享有博弈自由。相反,在算法应用场景中,当用户既无决定是否进行博弈之自由,亦不能参与博弈规则的制定和选取,只能被动承受博弈结果时,其处于完全丧失博弈自由的被支配地位,算法控制者和用户之间的横向权力关系由此形成。因此,本文认为,剥夺用户博弈自由是横向权力形成的实质标准,其以博弈锁定为具体的实现机制,须同时满足博弈退出封锁、博弈规则锁定和区分事由锁定三个条件。

1. 博弈退出封锁。博弈退出封锁即用户一旦进入算法系统,则丧失退出算法系统之自由。博弈退出封锁以用户进入算法系统、参与博弈为前提。算法控制者通过获取和分析个人信息开启博弈并非难事:一方面,个人或为享受算法红利、或为参与社会生活,会主动提供个人信息;另一方面,算法控制者为实现获利目的,也会极力攫取个人信息。由此可见,算法控制者设下的博弈从不缺少或积极参与、或被动入席的用户。

显然,若用户享有退出博弈的自由,则可避免处于算法控制之下。因此,博弈锁定首先以博弈退出封锁为必要,即用户处于算法控制之下而无退出权。判定用户是否遭遇退出封锁,主要以是否存在可替代的算法控制者为标准。可替代性则具体包括需求可替代性和供给可替代性两方面。需求可替代性,即是否存在可以提供相似产品或服务或其他算法控制者供用户选择;相似的算法控制者越少,用户退出特定算法系统后越难以获得相似产品或服务,特定算法控制者的需求可替代性也就越低。供给可替代性,则是基于其他算法控制者的供给能力来描述特定算法控制者的可替代性,具体应考量其他算法控制

者提供相似产品或服务所需的投入成本和收益;成本越高、收益越小,参与竞争的算法控制者越少,特定算法控制者的供给可替代性也就越低。特定算法控制者的需求可替代性或供给可替代性越低,用户退出算法系统后越难以获得相似的产品或服务,也就越易遭遇退出封锁。

退出封锁的典型情形,是算法控制者取得并滥用市场支配地位。正因如此,不少学者认为,算法价格歧视的形成以企业具备市场支配地位为必要^[6](P113)。然而,本文认为,算法控制者具有市场支配地位只是适用反垄断规制路径的必要条件,并非算法歧视的构成要件。当算法控制者虽不具有市场支配地位,但有强大的用户基础时,亦可形成退出封锁。用户基础包括用户数量、用户黏性和用户活跃度等。其原因在于,在网络效应影响下,用户基础往往决定算法控制者的需求可替代性和供给可替代性。一方面,算法控制者的用户基础越优,越能更好地满足用户需求,进而难以被其他算法控制者替代。如微信应用的用户数量巨大、用户黏性极高,极大程度地满足了用户的社交需求,已成为用户维系社会关系的必要工具。另一方面,拥有良好用户基础的算法控制者更易持续吸引和锁定更多用户,实现“赢者通吃”,其他算法控制者也就很难超越和替代。至于用户基础何时足以形成退出封锁效应,难以提出确切的量化标准,还须法官在司法裁判中综合考量用户数量、用户黏性、用户活跃度等要素后,再作判断。

2. 博弈规则锁定。算法权力的形成和实现还有赖于博弈规则锁定,即适用于用户的算法规则由算法控制者单方决定。一方面,算法决策依据的规则群(包括选取目标变量、设置分类标签、具体决策规则等)由算法控制者单方制定,用户无权参与;另一方面,算法控制者还掌握规则选取权,即用户适用于哪一类规则亦由其单方决定。对此,无法退出博弈的用户只能概括接受^①。相比之下,算法的博弈锁定远甚于格式合同对意思自由的限制。格式合同虽亦由规则提供者单方拟定,但相对人可概括接受也可概括拒绝,而遭遇博弈锁定的用户只能接受算法决策的结果,无权拒绝。

3. 区分事由锁定。区分事由锁定是算法权力形成的最关键要素,当用以区分用户的事由为用户固有的、难以改变的特性时,算法规则即完全锁定用户,用户不可避免地承受特定规则导向的决策结果。反之,若用户可以自由塑造与区分事由相关的特点,或区分事由乃用户可平等获得的特性时,用户可反向选取算法规则以脱离规则锁定。

区分事由锁定性的具体判定,须从区分事由是否具有人身依附性、是否具有可虚拟性两个方面进行考量。首先,就人身依附型区分事由而言,主要包含两类,一是反映个人与生俱来的人格特征的事由,如性别、种族、基因信息、年龄等;二是个人在日常生活和社会交往中积累形成的稳定特性,如资产状况、信用状况、社会地位等。这些区分事由依附于用户的人格或身份,难以重塑或消除,以此为基础进行区分决策往往会将用户锁定在与其所属分类相对应的规则群中。相反,若区分事由仅涉及用户偶然的行为、兴趣、偏好等,用户则可以通过获得、更改相应特性以加入或摆脱某一分类,避免不利决策。如算法控制者对新用户实行价格优惠时,由于新用户身份的获得取决于用户是否是首次实施购买行为,此种区分本质上为动态平等的区分而非固化的区分,老用户也曾为新用户,新用户也会成为老用户,通常不构成算法歧视^②。其次,即使区分事由的内容指向用户的固有特性,但可虚拟化,用户亦可通过虚构特点逃脱规则锁定(如女性用户在网络中使用虚拟的男性身份),故区分事由还须以不可虚拟化为必要。不可虚拟化通常存在于以下两种情形中,一是用户特性源于算法控制者自动收集和分析的数据,而非个人主动提供的信息;二是用户进入算法控制系统时须提供真实信息。

博弈退出封锁、博弈规则锁定和区分事由锁定最终导向博弈结果锁定:一方面,博弈结果因博弈规则自动执行必然发生,用户无力阻止;另一方面,用户因博弈退出封锁被迫承受既定的博弈结果,无论其

① 格式合同与算法歧视的效率实现机制亦有所区别,格式合同的效率体现在合同内容的重复适用,但算法歧视的效率恰恰体现在区分用户并实施差别对待上。

② 当然,对于在促销策略实施前即已成为老用户的主体而言,尽管其已被锁定在不能享受优惠的老用户群体内,但只要新用户之利益并非直接来源于对老用户的杀熟,其所遭遇的差别对待属于合理的差别对待,乃算法歧视构成之例外。

是否受到平等对待。

综上,算法歧视以横向权力关系存在为前提,横向权力关系的形成则由算法控制者实施博弈锁定、剥夺用户博弈自由塑造。算法控制者利用技术优势和信息鸿沟,将用户锁定于“必输”的博弈中,名义上的平等博弈不复存在,异化的算法权力关系由此形成。

三、算法歧视认定的具体要件及其例外

算法歧视尽管发生于新技术背景下,但仍保有传统歧视的根本特性。一方面,算法歧视同样需要权力支撑;另一方面,算法歧视本质上仍是对个人或群体的差别不利对待。因此,算法歧视的构成仍须具备传统歧视所要求的区分对待要件和差别不利后果要件。

(一)行为要件:算法区分决策

歧视皆因区分对待而起,如种族歧视是指“基于种族、肤色、世系、民族或人种的任何区别、排斥、限制或优惠”(详见《消除一切形式种族歧视国际公约》第一条);女性歧视是指“基于性别而作的任何区别、排斥或限制,其影响或其目的均足以妨碍或否认妇女不论已婚未婚而在男女平等的基础上认识、享有或行使在政治、经济、文化和社会、公民或任何其他方面的人权和基本自由”(详见《消除对妇女一切形式歧视公约》第一条)。算法歧视是依托于技术发生的歧视,同样以区分对待为要件,区分对待要件具体应从区分模式和区分事由两个方面考量。

1.以“区分—决策”或者“决策—区分”为运作模式。算法歧视首先以算法采取区分决策模式为要件,算法应用中未涉及区分决策的,即使对用户造成不利后果,也不构成算法歧视,而应交由其他法律制度调整。在算法歧视中,差别对待主要呈现为两种具体模式。一是“区分—决策”模式,二是“决策—区分”模式。“区分—决策”型差别对待是先进行用户分类,再分别决策,区分性直接表现在决策依据上,是直接的差别对待,其造成的歧视为直接歧视。“决策—区分”型差别对待是隐蔽的差别对待,算法规则表面上平等地适用于每个用户,但其结果上造成了对不同群体之间的差别对待。“决策—区分”型差别对待属于间接差别对待,其所造成的歧视属于间接歧视^①。

个性化决策模式同样属于区分决策模式,所谓的“个性化”,只不过是信息不对称和算法隐蔽性的双重作用下,用于“包装”区分决策的粉饰之辞,实际上仍是通过分类模型实施的精致差别对待。不同的是,在个性化模式下,算法区分事由及由此划分出的群体类型复杂精细且保持动态,算法控制者将用户划入微分类中,具体决策也在对应的微分类中灵活变动,由此营造出用户被当作独特个体对待的假象。

此外,分类正确与否并不影响歧视的构成。算法区分决策模式并不追求分类的真实性和准确性,相反,其是以分类的高容错率为代价,换取决策的高效简化。错误识别用户特点并将其划入错误群体的情形,在算法决策中并不鲜见,如将居住于贫困地区的富人识别为穷人^{[7](P9)}。区分的正确与否并不影响歧视的成立,用户被错误区分并因此遭受不合理的差别对待,亦构成歧视。

2.突破“区分事由法定”。在传统歧视中,区分事由主要包括性别、种族、民族、宗教信仰等。传统反歧视规制要求区分事由必须是受法律保护的个人特征,个人因法定事由以外的特征受到差别对待的,则很难被认定为歧视^{[8](P775)}。然而,算法歧视的认定应当摒弃“区分事由法定”的限制。原因在于,伪中立性的代理标签(如以身高、体重代替性别,以清真食品的消费记录代替宗教,以邮政编码代替民族),以及算法自动区分决策背后存在大量潜在、隐蔽且难以预知的区分事由,“区分事由法定”将导致诸多算法歧视现象游离于法律规制之外。因此,有必要突破“区分事由法定”的限制,保持区分事由范畴的灵活性

^① 间接歧视即一项法律、规则、标准尽管表面中立,但最终造成特定群体或个人遭遇差别性不利对待。间接歧视类似于差别性影响类歧视,前者是英国法中与直接歧视并列的歧视类型,后者是美国法中与差别性对待并列的歧视类型,二者的主要区别体现在是否以主观故意为要件。

和开放性^①。

(二) 结果要件: 差别性不利后果

算法歧视的具体认定, 还以用户遭受差别不利对待为结果要件: 一方面, 算法区分决策的结果须同时具有不利性和差别性; 另一方面, 此种差别不利后果无须达到社会排斥程度。

1. 区分对待结果的不利性。区分对待结果的不利性即“利益的应得而未得”, 表现为用户应当获得的利益被剥夺或克减, 前者如用户应当取得信贷资格而未取得, 后者如用户获批的信贷额度低于其应当取得的获批额度。

值得注意的是, 不利后果不限于现实利益分配的应得而未得。在评分算法中, 不利评分尽管不属于直接的利益分配, 但若以评分结果为资源配置的直接依据, 则评分的不利与利益分配的不利具有一致性, 前者亦应被纳入不利后果的范畴。

2. 区分对待结果的差别性。区分对待结果的差别性是指, 不利后果必须是相对于用户的“应得”而言的, “不利”是“相对的不利”。差别性要件要求用户主张受到算法歧视时, 证明不利后果的差别性。提供对照是最直接的证明方法, 用户提供满足以下条件的对照即可证明其受到差别的不利对待: (1) 在所涉区分事由上与他人有不同特征; (2) 在其他与决策结果相关的事由上, 与他人相似或者一致; (3) 他人取得更有利的分配结果。然而, 算法歧视中, 由于个性化策略、复杂隐蔽的区分事由、用户之间存在的信息阻隔等多种原因, 用户寻找对照的难度大, 面临举证难题。因此, 有必要扩充对照范围。一种可行的方案是, 从横向对照向纵向对照扩充, 即可以用户自身过往经历作为对照, 证明其因某项事由发生改变, 而遭受差别不利对待。纵向对照在我国传统反歧视实践中已有应用, 如在“卢某与深圳市兆新能源股份有限公司一般人格权纠纷案”(广东省深圳市罗湖区人民法院[2020]粤0303民初26265号民事判决书)中, 原告主张其本已通过入职面试, 但被告在得知其携带某种病毒后又拒绝录用, 构成歧视, 法院对此予以支持。另一种方案则是承认假想对照, 即原告无须提供真实存在的对照, 只须证明其一旦消除或取得与区分事由相关的某种特性, 便有可能获得更有利的对待即可。对于用户主张的差别不利对待, 算法控制者可通过证明不存在差别不利对待, 或差别不利对待具有合理性, 而免于算法歧视规制。

3. 差别不利对待无须达到社会排斥程度。还有观点认为, 算法歧视造成的不利后果以达到社会排斥程度为必要^[9](P371)。本文对此持反对观点。社会排斥是指弱势群体被部分或全部排斥在充分的社会参与之外, 难以获得基本的生存发展资源。反歧视是保护弱势群体不因某种难以改变的特征而丧失基本的生存发展权利。造成社会排斥的差别对待当然构成歧视, 但歧视本身不应以不利性达到社会排斥程度为要件。原因在于, 社会排斥程度门槛过高, 过度限缩了算法歧视的规制范畴。若引入此要件, 则只有造成机会剥夺或严重利益克减的差别对待才能被认定为歧视。然而, 在算法区分决策应用领域, 最普遍的差别对待往往在结果上差异不大, 算法控制者正是通过将结果差异维持在较低水平, 悄然获取巨大利润。可见, 规模化的轻度差别对待才是算法正义的主要破坏力量, 也是反算法歧视的主要规制对象。

此外, 需要强调的是, 算法歧视的不利后果与侵权责任中的损害后果存在区别。首先, 歧视是利益分配中的应得而未得, 是“做加法”过程中利益的未获得或未充分获得, 是对照下的相对的不利; 而损害后果则是个人权利受到侵害, 是对固有权益“做减法”, 是绝对的不利。其次, 反歧视本质上是群体利益和社会利益的保护机制。反歧视所保护的总是具有某种群体性特征, 反个体歧视最终是为了反群体歧视; 而侵权责任具有更鲜明的个人主义, 其关注的是个人权利是否受到侵害, 个人是否具有某种群体特征不在其考量范围内^[10](P1119-1120)。

^① 事实上, 我国司法实践基于维护分配正义之现实需要, 早已突破传统歧视事由的限制, 地域歧视、基因歧视等均受到规制。具体案例可以参考浙江省杭州市中级人民法院(2020)浙01民终736号民事判决书; 佛山市中级人民法院(2010)佛中法行终字第381号判决书。

(三) 算法歧视的构成例外:合理的差别对待

在博弈锁定形成后,算法控制者对用户进行区分对待,强迫其承受差别性的不利后果,即构成算法歧视。然而,差别对待具有合理性的,例外地不构成算法歧视。差别对待合理性的证明义务由算法控制者承担。合理性证明主要涉及三方面:一是差别对待目的的合理性,二是差别对待目的与区分事由之间的合理关联性,三是差别对待的合比例性。

1. 差别对待的目的合理性。算法区分决策的目的可归结为追求效率、实现获利,有多种具体表现形式,如个性化广告推送意在精准吸引潜在消费者,个性化定价意在扩充消费者群体、攫取消费者剩余,贷款信用评分意在降低坏账风险等。然而,追求效率并不当然具有合理性,尤其是在用户被剥夺博弈自由时,算法控制者极可能为提升效率,将自身利益凌驾于用户利益之上。因此,目的合理性要求决策目的实现并未侵害用户利益,若算法控制者的决策目的的实现以侵害用户利益为前提,差别对待目的不具有合理性。

2. 差别对待目的与区分事由间的合理关联性。合理关联是指具有正当且必要的关联。如信贷资格的分配与用户的履约程度(如是否依约定偿还贷款)之间具有合理关联,但与用户的消费习惯则无必要关联,尽管后者可能反映用户的经济水平和消费理性程度;又如差别定价与用户特殊需求往往具有必要正当关联,但与用户的价格敏感度则无必要关联,后者并非实质的交易条件,算法控制者以后者作为差别定价的依据,不符合正当的交易习惯。

合理关联的证明首先要求算法控制者揭示真实的区分事由。在算法控制者使用代理标签进行区分决策的情形下,由于代理标签与决策目的之间关联较为疏远,算法控制者须主动揭示真实的区分事由,才能完成合理关联的证明。可见,合理关联证明义务客观上有利于解决真实区分事由的认定问题。合理关联性还要求算法控制者证明不存在其他可替代的、能够防止或者缓解不利后果的区分事由。算法控制者主张区分事由不可替代时,用户可提出可替代性事由或其他相反证明。

3. 差别对待的合比例性。差别对待合比例性证明具体包括两个层面:一是差别对待的区分事由与不利后果之间具有合比例性;二是差别对待的不利后果与决策目的之间具有合比例性。

区分事由与不利后果之间的合比例性,要求算法控制者在复合的区分事由中,为各区分事由设置合理权重。区分事由在决策目的实现中的作用程度,应与其所生后果的差别性和不利性程度相匹配,次要的区分事由不应主导差别对待。此一证明义务,要求算法控制者阐明特定区分事由在决策目的实现中的具体作用和所占权重,并证明二者具有合比例性。以某讯的征信评分算法为例,其以履约指数(如是否如期偿还信用卡借款)、安全指数(如个人信息是否准确、账户安全性是否足够高)、财富指数(如个人资产状况)、消费指数(如消费偏好、消费频率)以及社交指数(如用户的人脉关系)作为评价指标,对用户进行信用评分^[1]。算法控制者必须揭示各指数在征信评分中的作用及其权重分布,并说明权重分布的合比例性。显然,履约指数与信用评分的关联最为紧密,其在信用评估中应占最高权重,而消费指数、社交指数仅与信用存在间接的、不稳定的关联,权重设置应当处于低位。若该评分算法将消费指数和社交指数作为主要的评分指标,则存在区分事由与差别不利后果的比例失当,可能构成对消费频率较低者或者社交偏保守者的歧视。

差别对待的合比例性还包括不利后果与决策目的之间的合比例性,此种合比例性旨在防止算法控制者过分规避商业风险,损害用户利益。为实现效益最大化,算法控制者可能倾向于利用信息优势,一边摒弃风险用户,一边从“合格”用户处极力攫取利益以作弥补。在此种利益补偿机制中,双边用户都沦为算法控制者的牺牲品。不利后果与决策目的间的合比例性,意在规制算法控制者恣意的趋利避害,要求算法控制者在实现商业目的时,尽可能减轻针对用户的不利对待。如在利用信用评分算法进行贷款审批时,算法控制者应灵活配置贷款资格、贷款额度、增信措施、偿还期限等要素,以适用于信用评分不同的用户,而不能为了规避风险,一概剥夺信用评分偏低群体的贷款资格。不利后果的合比例性还要求

算法控制者采取措施缓解不利后果,如优化算法模型、为用户提供反馈渠道等。算法控制者主张不存在消除或缓解不利后果的措施的,用户可以提出相反证明。

须明确的是,本文认为,算法歧视不以算法控制者是否具有主观过错为要件,原因主要有两点:一是算法决策的隐蔽性特征,导致过错认定客观上变得十分困难;二是算法控制者为算法歧视风险的制造者、控制者和获益者,理应分担歧视风险。从域外立法来看,英国的反歧视规制不以主观过错为要件^[12](P619);而美国反歧视制度也明显出现主观要件式微的趋势,一方面,受到规制的歧视类型从差别对待扩张至差别性影响,后者不再以主观故意为要件;另一方面,差别对待类型中,主观故意要件的认定也趋于客观化^[9](P372-373)。

至此,算法歧视以横向权力关系存在为前提,并须同时满足行为要件和结果要件。三个要件及其相互关系简要总结于图1。

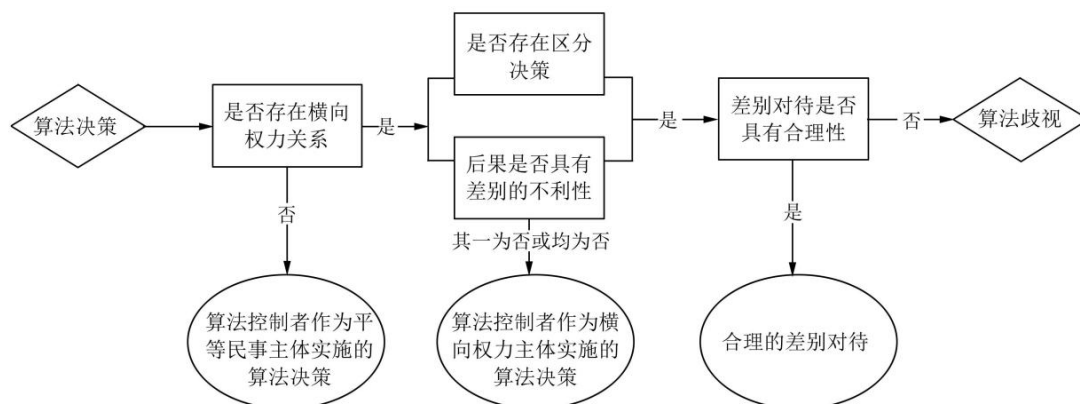


图1 算法歧视的认定过程示意图

结合图1,笔者还须作两点说明:一是,横向权力关系为算法歧视发生的首要要件,若不存在横向权力关系,即使用户遭受差别不利对待,也不构成算法歧视。二是,在横向权力关系形成后,若不满足区分决策要件抑或差别不利后果要件,亦不构成算法歧视,但算法控制本身是否违法则未有定论,当然,典型的违法情形除外,如滥用算法权力侵犯言论自由、实施欺诈等。

四、常见算法决策类型中的算法歧视认定

在有关算法歧视的讨论中,常被列举的算法决策类型包括:单纯呈现歧视性内容的搜索算法或人机对话、“大数据杀熟”、评分算法,下文将结合构成要件对这几种算法决策是否构成算法歧视进行分析。

(一) 类型一:搜索算法和人机对话对歧视的镜像呈现

有学者将搜索算法单纯呈现已有偏见和歧视,以及人机对话过程中人工智能“说出”歧视性言辞等现象认定为算法歧视。本文认为,这是典型的算法歧视认定泛化^[13](P141)。应当认为,上述两种算法决策只是对已有歧视和偏见的镜像呈现^①,其因欠缺区分决策的行为要件,不构成算法歧视。从算法应用模式来看,搜索结果的产生由爬虫技术、索引技术和查询展现技术完成,人机对话内容则是通过用户输入、自然语言理解、对话管理和自然语言的生成与输出产生,二者均未涉及区分模型的应用,不存在对用户的区分,更不存在区分后的差别对待。搜索结果、对话内容中的歧视内容源于污染数据,而数据中的偏见又可追溯至社会现实中的已有偏见。例如,某国执法机关受社会种族偏见的影响,更倾向于逮捕有

① “镜像”通俗意义上是指“像照镜子一样”,本文所述的“镜像呈现”是指算法仅单纯展示和呈现社会中已经存在的偏见和歧视,而并未通过决策制造新的歧视的情形。

色人种,算法基于上述社会现实得出“有色人种的犯罪风险更高”的分析结果,只是对该国社会种族偏见的镜像呈现,不构成算法歧视。然而,算法在镜像呈现社会偏见和歧视的基础上作出新的歧视性决策,则另当别论。例如,若该国法院依“有色人种的犯罪风险更高”这一算法分析结果,对犯罪嫌疑人中的有色人种判处更长刑期,就构成算法歧视。

当然,算法呈现歧视内容时并非不受其他法律规制。依据我国《互联网信息服务管理办法》第15条的规定,互联网信息服务提供者不得制作、复制、发布、传播反对宪法所确定的基本原则的信息,煽动民族歧视的信息,侮辱或者诽谤他人的信息,侵害他人合法权益的信息,以及含有法律、行政法规禁止的其他内容的信息。若算法呈现的歧视内容符合上述情形,则算法控制者须依《互联网信息服务管理办法》第20条承担法律责任。此外,《互联网信息服务算法推荐管理规定》(下文简称《算法推荐管理规定》)第六条规定算法推荐服务提供者“不得利用算法推荐服务传播法律、行政法规禁止的信息,应当采取措施防范和抵制传播不良信息”,第九条要求算法控制者承担信息内容管理义务,对违法信息采取消除等处置措施。依此规定,算法控制者传播歧视性信息或未依法履行信息审查义务的,亦可能面临法律责任。

除上述两种情形外,因算法决策之外的原因造成的算法红利分配失衡,同样也因不满足区分决策要件,不构成算法歧视。算法技术和应用在地区之间(如经济发达地区和经济欠发达地区)、群体之间(如不同学历群体、不同年龄群体之间)普及不平衡导致的算法红利分配失衡,通常应归因于社会经济发展水平差异、收入差异、群体使用智能设备的能力差异等客观因素,其并非由算法决策导致,不应由算法控制者担责。解决算法红利分配失衡,有赖于系统性的制度合力,以推动经济平衡发展、加快技术普及、促进教育发展、完善社会保障体系等为主要应对策略。

(二) 类型二:“大数据杀熟”

“大数据杀熟”是目前最受争议的算法决策类型,其最初是指对具有购买经历的消费者(即“熟客”)采取更高定价的个性化定价策略。随着个性化定价算法的进步,杀熟的范围从“熟客”扩展到被算法精准捕捉画像的“熟人”,杀熟程度趋近于完全价格歧视。不少学者认为,“大数据杀熟”攫取消费者剩余、侵蚀社会公平正义^[14](P27),应对其进行强力规制。《深圳经济特区数据条例》第69条禁止利用数据分析对交易条件相同的交易相对人实施不合理的差别待遇,正式对“大数据杀熟”亮剑。之后出台的《个人信息保护法》第24条亦禁止自动化决策中的不合理差别对待。也有学者对“大数据杀熟”持“一事一议”的谨慎态度,认为“大数据杀熟”有增加社会福利、优化资源配置、促进商业竞争的积极效果,不能一概冠以歧视的污名^[15](P85-98)。应当认为,“大数据杀熟”是否构成算法歧视不可一概而论,应作个案分析。

1. 是否存在博弈锁定。传统市场中的杀熟宰客并不鲜见,消费者通常“自认倒霉”,法律对此也有较高的容忍度。然而,一旦转换到算法场景中,杀熟则迅速引发民众担忧和立法关注。此间态度截然的根本原因在于,算法中的杀熟是建立在剥夺消费者博弈自由的基础上的:传统市场中,商家与消费者基于各自掌握的有限信息,就是否开展交易以及交易价格进行平等磋商,杀熟是自由博弈的结果,具有偶然性;算法场景下,自由博弈很可能遭到根本破坏,消费者完全丧失议价空间和议价能力,难以逃脱被杀熟的宿命。可见,“大数据杀熟”是否构成算法歧视,首先须判定消费者是否遭受博弈锁定。为具化分析,本文以某宝和某滴两种常用平台为例展开论述。

在某宝平台上,普通商家通常很难对消费者实施退出封锁。稍有网购经验的消费者即可通过简单的消费策略从平台充分获取有效的价格信息,进而自主选取有利定价。即使部分消费者因信任特定商家而遭遇杀熟,亦属自由博弈之结果,产生的不利后果应由当事人和市场自行消化,不属于反歧视的规制范畴。然而,当商家存在价格合谋时,则可能发生博弈锁定。如同一行业所有商家采取同一定价算法,消费者与任一商家交易,都遭遇同样的定价策略;又如各商家借助算法实时追踪并效仿其他竞争对手的价格策略,并达成提价默契,对消费者实施默许共谋封锁^[16](P51)。某滴平台对乘客用户的博弈锁定则十分明显。一是某滴平台拥有庞大的车主用户,乘客用户对平台依赖程度高;二是某滴平台掌握了

派单权和定价权,为乘客用户选取交易对象,为每一笔交易定价,甚至在订单结束时自动完成价款扣划,乘客和司机均无议价权,乘客只能被动接受某滴平台的价格策略。

此外,还须考量区分事由的锁定性。定价算法模型用以区分消费者的事由是否具有锁定性,应视情况而定。例如,用户的支付能力取决于用户的资产状况、职业状况,是个人难以改变的特性,具有明显的锁定效应,以支付能力作为差别定价的依据,要求支付能力更强者支付更高价款,往往构成算法歧视;又如,某滴平台在出行早高峰、节假日等用车高峰期调高价格,是以市场供求变动作为定价基础,既符合市场规律,又未形成博弈锁定,乘客不能因其在早高峰支付更高车费而主张其遭受歧视。

2. 是否存在差别性不利后果。“大数据杀熟”建立在消费者分类基础上,是典型的区分决策模型,其是否构成歧视,还须进一步判定区分结果是否存在差别性和不利性。最常见的杀熟为消费者在交易中因“熟客”身份反遭更高要价,此种差别对待显然具有差别性(与“生客”相比)和不利性(出价更高)。如北京市消费者协会两名体验人员发现同一房间“熟客”需多支付9元^[17];又如,在“‘大数据杀熟’第一案”中,原告作为平台会员,在预订房间时反被要求支付远高于挂牌价(1300元)的“会员价”(2889元)^[18]。

3. 差别定价是否具有合理性。若算法控制者证明其差别定价具有合理性,则不构成算法歧视。差别定价的合理性证明中,差别定价的目的合理性、差别定价与区分事由之间的合理关联性为最关键的两项要素。目的合理性须判断算法控制者的商业目的是否侵犯了消费者的利益。通常而言,差别定价的商业目的在于增加利润、吸引更多消费者,其目的实现有利于对企业形成经济激励、增加社会整体福利,因而具有一定的正当性。然而,当差别定价所增加的社会福利源于对特定消费者群体(如价格敏感度低的消费者群体、支付能力强的消费者群体)的利益攫取时,即使社会整体福利有所上升,差别定价也不具有合理性。反之,当社会整体福利的增加来源于商家自身的利益让渡时,差别定价应被视为正当的商业策略,具有合理性。差别定价目的与区分事由之间的合理关联性,即要求差别定价与区分事由之间具有直接、必要的关联。应当认为,当用户存在交易条件上的实质性差别时,差别定价具有合理性,实质性差别包括用户在交易安全、交易成本、信用状况、交易环节、交易持续时间等方面的差别,以及其他符合正当交易习惯和行业惯例的交易条件上的差别,如交易相对人的特殊需求。

(三) 类型三: 评分算法

评分算法主要指算法控制者基于特定商业目的,对用户的某种特性进行量化评估的算法应用。评分算法是典型的区分决策,其是否构成歧视同样需要进行个案分析。下文以产品、服务质量评分算法和信用评分算法为例,作简要论述。

1. 产品、服务质量评分算法。实践中,平台会为消费者设置评价渠道,供消费者对产品、服务及产品销售者、服务提供者进行评分,评分结果则直接影响产品销售者、服务提供者未来的交易机会。通常情况下,产品、服务质量评分算法不会对平台上的产品销售者、服务提供者造成歧视,原因在于,此种算法的评分指标是与产品或服务质量直接相关的合理因素,平台只确定评分指标,无权决定评分结果,因而不存在博弈锁定和算法控制。如某滴平台车主的评分指标包括车内环境、车主驾驶技术、路线熟悉度等,某宝卖家的评分指标包括商品描述一致性、客服态度、商品运输速度等,车主和卖家可通过改善其服务或产品等方式增加评分。当然,实践中也存在例外情形。以打车软件某步为例,某步的车主用户因属于有色人种而遭遇乘客用户的不当评分,工作机会和收益也因此减少。此时,某步的车主评分机制已经潜在包含了种族这一锁定性区分事由,某步作为算法控制者应采取措施予以消除,否则构成算法种族歧视^[19]。

2. 信用评分算法。信用评分算法是信用评估机构基于个人基本信息、银行信用信息、消费信息、资产信息等各项数据指标对用户信用进行评分的技术应用。信用评估机构可能出于不同目的实施信用评分,如银行的信用评分目的可能是分配信用贷款资格、确定贷款额度,芝麻信用、腾讯征信的信用评分目的则可能是确定和调整其服务内容和服务质量(如是否向用户提供免押金租赁服务、先用后付服务等)。

信用评分算法多应用在消费、住房、教育等关键的社会参与领域,歧视风险较高。当用户遭遇偏低的信用评分时,是否构成算法歧视,还须具体分析,其中的关键要件主要涉及是否存在博弈锁定、是否造成差别不利后果。首先,信用评分的具体结果通常由信用评估机构依托已有数据和算法规则直接作出,且评分指标是用户难以改变的身份、行为特征、资产状况、历史信用状况等,故信用评分的主导权由信用评估机构掌握。但信用评估机构是否构成博弈锁定,还取决于用户能否拒绝信用评分服务。当用户必须接受信用评分时,则发生博弈锁定,如个人申请住房贷款必须有合格的信用评估;当用户可拒绝信用评分时,则不存在博弈锁定,如用户有权拒绝芝麻信用评分,且拒绝后支付宝使用不受实质影响。其次,信用评分不当偏低是否会造成差别性不利后果,也须依场景而定:在某些场景中,信用评分未造成不利后果,如评分尽管不当偏低但仍与用户应得的评分处于同一分数区间,不影响决策结果;在另一些场景中,信用评分不当偏低则会产生差别的不利后果,如银行信用评分不当偏低,导致个人不能取得住房贷款资格。当然,无论是否构成算法歧视,对于错误的信用评分,用户均可依据《个人信息保护法》第46条行使个人信息更正权,要求算法控制者更正评分。

甄别私法关系中的算法歧视,仅仅是反算法歧视的第一步。在完成算法歧视认定的基础上,进一步构建算法歧视规制的体系化规则,同样是立法和司法的重要课题。由于我国尚不存在完备的反歧视私法规则,宜将一般人格权侵权规制作为算法歧视规制的权宜之计:算法控制者实施算法歧视的,可认定为算法控制者侵害用户的一般人格权(人格平等)。侵权的具体判定适用算法歧视的特殊构成要件,损害赔偿的计算则以“应得而未得”部分为标准,如用户因遭遇歧视性定价而支付多余价款的,算法控制者应返还用户多支付的部分;算法歧视造成用户严重精神损害的,算法控制者还应承担精神损害赔偿责任。当然,算法歧视规制还可通过反垄断规制、个人信息保护、消费者权益保护等路径实现。除此之外,行政机关也应强化算法监管。一方面,应加强对算法源代码和算法分析数据的审查,以防止算法控制者在算法系统中植入偏见,及时排除分析数据中的污染数据;另一方面,应引导企业有序、适当地公开其算法,以推动算法透明,防范算法滥用风险。事实上,与其他算法不正义问题一样,算法歧视的症结在于算法权力。因此,反算法歧视的规制重点应为防止算法权力的形成和滥用。对此,我国有关立法已有积极动作,如《电子商务法》第18条规定电子商务经营者向消费者提供商品或服务的搜索结果时,应同时提供不针对消费者个人特征的选项;《算法推荐管理规定》第17条更是赋予用户关闭算法推荐服务的权利和选择、删除用户标签的权利。这些规定均有利于削弱算法权力、消除算法控制,从根源上减少算法歧视的发生。

参考文献

- [1] 崔靖梓.算法歧视挑战下平等权保护的危机与应对.法律科学(西北政法大学学报),2019,(3).
- [2] Ignacio N.Cofone. Algorithmic Discrimination Is an Information Problem. *Hastings Law Journal*, 2019, 70(6).
- [3] 潘芳芳.算法歧视的民事责任形态.华东政法大学学报,2021,(5).
- [4] 王德夫.论人工智能算法的法律属性与治理进路.武汉大学学报(哲学社会科学版),2021,(5).
- [5] 马克斯·韦伯.经济与社会:下卷.林荣远译.北京:商务印书馆,1997.
- [6] 周围.人工智能时代个性化定价算法的反垄断法规制.武汉大学学报(哲学社会科学版),2021,(1).
- [7] 凯西·奥尼尔.算法霸权.马青玲译.北京:中信出版集团,2018.
- [8] Kenji Yoshino. The New Equal Protection. *Harvard Law Review*, 2011, 124(3).
- [9] Denise G. Reaume. Harm and Fault in Discrimination Law: The Transition from Intentional to Adverse Effect Discrimination. *Theoretical Inquiries in Law*, 2001, 2(1).
- [10] Sandra F. Sperino. Let's Pretend Discrimination Is a Tort. *Ohio State Law Journal*, 2014, 75(6).

- [11] 腾讯信用是什么? 腾讯网, 2020-09-02. [2021-08-11]<https://support.qq.com/products/7438/faqs/1366>.
- [12] 谢增毅. 美英两国就业歧视构成要件比较——兼论反就业歧视法发展趋势及我国立法选择. 中外法学, 2008, (4).
- [13] 闫坤如. 人工智能的算法偏差及其规避. 江海学刊, 2020, (5).
- [14] 马长山. 智能互联网时代的法律变革. 法学研究, 2018, (4).
- [15] 喻玲. 算法消费者价格歧视反垄断法属性的误读及辨明. 法学, 2020, (9).
- [16] 阿里尔·扎拉奇, 莫里斯·E. 斯图克. 算法的陷阱: 超级平台、算法垄断与场景欺骗. 余潇译. 北京: 中信出版社, 2018.
- [17] 王薇, 赵婷婷. 北京市消协发布“大数据杀熟”问题调查结果. 北京青年报, 2019-03-28. [2021-10-01]http://epaper.yynet.com/html/2019-03/28/content_323364.htm?div=-1.
- [18] 诸未静, 李紫瑄. 携程否认“大数据杀熟”背后: 事实性认定的难点在哪里? 21 世纪经济报道, 2021-07-19. [2021-10-01]<https://view.inews.qq.com/a/20210719A05TLG00>.
- [19] Josh Eidelson, Bloomberg. Uber Sued by Non-white Drivers Alleging Race-biased Ratings Lead to Firings. Fortune, 2020-10-27. [2021-10-07]<https://fortune.com/2020/10/26/uber-lawsuit-drivers-fired-low-ratings-racial-discrimination>.

Determination of Algorithmic Discrimination

Ning Yuan (Wuhan University)

Abstract Existing research on anti-algorithmic discrimination lacks the logical premise of discrimination determination, which leads to the generalization of algorithmic discrimination and needs to be corrected. The contradiction between algorithmic application and algorithmic justice has been over-emphasized, which requires a clear determination rule for algorithmic discrimination. Horizontal power relations are a necessary field for algorithmic discrimination: algorithmic discrimination only exists when there is horizontal power. The horizontal power relationship is formed when the algorithm controller classifies users with locked distinction, unilaterally decides the algorithm rules applicable to users, and users can only passively accept the decision results without the freedom to exit the algorithm system. In terms of specific elements, algorithmic discrimination is based on the act of differentiation, and on the result of differentiation and disadvantage. Algorithmic controllers are exempt from anti-discrimination regulation by justifying differential treatment. The pure presentation of social discrimination in search algorithm and human-machine dialog algorithm as well as algorithm dividend distribution imbalance caused by social-economic development and imbalanced application of algorithm does no lead to algorithmic discrimination. Differential treatment caused by profiteer on big data and scoring algorithm does not necessarily constitute discrimination and should be analyzed case by case.

Key words algorithmic control; algorithmic discrimination; horizontal power; distinguishing decision; differential adverse consequence; profiteer on big data; scoring algorithm

■ 收稿日期 2021-11-04

■ 作者简介 宁 园, 法学博士, 武汉大学法学院特聘副研究员; 湖北 武汉 430072。

■ 责任编辑 李 媛