数据表征的法律治理*

朱　悦**

摘要：数据表征将原始数据转化为人工智能模型能够高效处理的信息，是连接容易治理的数据和难以治理的模型的桥梁。国内外法律正在探索通过治理数据表征治理人工智能：国内针对特定场景算法的专门立法明确纳入数据表征，但难以具体适用；国外个人信息保护执法对法律如何适用于其表征有具体说理，但尚未形成体系立法。综之，通过适当借鉴国外说理，国内既有法律可以清晰适用于数据表征，从而增强人工智能治理。结合法律、市场、伦理、架构等层面的治理进展，数据表征有望成为人工智能治理——包括未来相应立法——中的关键环节。

关键词：数据表征　人脸识别　推荐算法　人工智能治理　《个人信息保护法》

数据表征将原始数据转化为人工智能模型能够高效处理的信息，其发展也是近年来人工智能技术，包括深度学习技术迅速发展的重要因素（Bengio，Courville & Vincent，2013；LeCun，Bengio & Hinton，2015），数据表征因而构成连接原始数据和模型的桥梁。从人工智能治理特别是其法律治理的角度看，原始数据的治理相对容易。《中华人民共和国个人信息保护法》（以下简称《个人信息保护法》）等法律法规都已对原始数据有所规范。模型在技术上更加复杂，治理也相对困难，相关法律法规很少涉及人工智能模型，即使涉及通常也只有原则性的、不适用的规则。例如《个人信息保护法》第二十四条要求“利用个人信息进行自动化决策，应当保证决策的透明度和结果公平、公正”，但并未包括更具体的内容。法律治理由此开始关注可以桥接数据与模型的表征，体现为国内外都开始出现涵盖数据表征的立法和执法。具体来说，国内主要通过立法进行治理，例如《互联网信息服务算法推荐管理规定》（以下简称《管理规定》）第九条

* 本文受益于与周辉、许可和王睿的相关讨论，特此致谢。

** 朱悦，中国社会科学院文化法制研究中心。

和第十条针对属于数据表征的特征、标签和兴趣点设定义务;国外更多是在具体执法中针对数据表征展开分析,例如英国和澳大利亚数据当局在针对人脸识别服务提供商 Clearview AI 的调查中,对属于数据表征的人脸矢量展开分析(Australian Information Commissioner and Privacy Commissioner [OAIC],2021)。相应地,国内法律仍然缺乏对数据表征的具体适用,导致难以实际实施;国外也鲜有涵盖数据表征的立法体系,从而失之过度零碎。于是,尽管数据表征在人工智能的新近发展中非常重要,但其法律治理的重要性和潜力却尚未充分体现,国内外相关的法律治理探索也有待梳理、整合。

有鉴于此,本文第一节在介绍数据表征和实践的基础上,进一步阐明通过治理数据表征推进人工智能治理的重要性和潜力;第二、三节分别梳理国内外法律治理数据表征的尝试,国内主要是不同位阶的立法,国外更多的是不同法域的执法;第四节加以整合,以国内法律为出发点,阐释如何使现有法律清晰适用于数据表征;第五节首先分析推进数据表征的法律治理如何促进人工智能治理,然后探讨法律以外通过规范、市场和架构(Lessig,2006)等方式治理数据表征的进展如何促进人工智能治理,之后展望数据表征及其治理如何作为未来人工智能治理立法中的关键环节。

一、数据表征与人工智能治理面临的挑战

近年来,数据表征将原始数据转化为模型处理数据的方式经历了显著的变化,数据表征在显著提高处理效率的同时,其可理解性也在显著下降。这有三方面影响:一是聚焦数据表征的表征学习成为推动人工智能发展的重要方向;二是可理解性的下降与透明性、可解释性等人工智能治理问题密切相关;三是数据表征技术的变化引起了概念的变化,需要明确人工智能实践中许多常见的概念实际上都是表征的不同说法。数据表征还与作为人工智能前沿发展趋势的“大模型”[①](Bommasani,Hudson & Adeli et al.,2021)和开源模型社区存在密切关系,于是亦有必要简述这些趋势对法律治理可能产生的影响。

数据表征的显著变化指人工智能模型本身,而非人类专家主导了原始数据

① foundation model 在国内存在多种译法,除“大模型”外,还有“基础模型”“通用模型”等,本文采用比较常见的“大模型”译法。

的转化过程(Bengio, Courville & Vincent, 2013; 周志华, 2016)。简言之,反映学习对象本身属性或表现的原始数据,未必适合直接作为人工智能模型的输入。因此,在很长一段时间内,需要人类专家通过理解、清洗、转换、选择和探索性分析等方法,在原始数据的基础上找到最适合模型的表征,这一过程通常被称为"特征工程"①(厄兹代米尔、苏萨拉, 2019)。随着人工智能技术,特别是深度学习技术的发展,模型本身即可实现将"初始的、与输出目标之间联系不太密切的输入表示,转化成与输出目标联系的表示"②(周志华, 2016),从而替代了人类专家的角色。由此,自动化地学习最适合模型的数据表征这一需求催生了被称为"表征学习"的专门方向(Bengio, Courville & Vincent, 2013)。不仅是技术本身主导了数据表征的过程,模型学习数据表征的进步也推动了技术的发展。典例是从自然语言处理开始,并很快席卷其他场景的一系列人工智能模型,包括转换器模型、来自转换器的双向解码表征模型和稳健优化的来自转换器的双向解码模型③,等等(Vaswani, Shazeer & Parmar et al., 2017; Devlin, Chang & Lee et al., 2019; Liu, Ott & Goyal et al., 2019)。这些模型在更有效地表征语义信息的同时,相应实现了低计算成本、高可扩展和通用任务等良好的性能。正是这些模型——特别是 OpenAI 基于转换器模型实现的、能够高效合成各类文本的 GPT-2 和 GPT-3 模型(Radford, Jeffrey & Child et al., 2019; Brown, Mann & Ryder et al., 2020),使得算法治理的许多担忧愈发趋于现实。其中,OpenAI 及其合作方在发布 GPT-2 模型时已经预见了算法偏见、深度伪造、仇恨言论和有关的新型安全威胁等挑战(Solaiman, Brundage & Clark et al., 2019),但这些挑战至今也很难被认为已然完全解决。数据表征由模型实施这一变化,是人工智能技术发展的里程碑,也是人工智能治理挑战的发轫地。

数据表征的显著变化不仅通过促进技术发展形成一般性的治理挑战,其自身也直接产生了更加具体的治理挑战。在由人类专家实施特征工程的时代,这些专家的第一步便是理解原始数据(厄兹代米尔、苏萨拉, 2019)。之后通过转换、选择等方式生成的数据表征,同样建立在理解的基础上。于是,作为模型输入的数据及其表征始终没有脱离人——至少是人类专家——的理解范围,人也

① "表征"和"特征"是相近的概念,"表征工程"这一概念同样存在,但远不如"特征工程"常用。

② "表征"与"表示"是 representation 的不同译法。

③ 这三个模型更常见也更方便的表述分别是 Transformer、BERT 和 RoBERTa。

就始终处在人工智能的环内。随着模型接管学习数据表征的任务,数据表征不再想当然地具备对人而言的可理解性。实践中,这些表征总是一些看似没有意义的数字。对于发生显著变化以后的数据表征,其是否可能,又应如何对人解释、为人理解,成为亟待解决的前沿问题。然而,针对具体场景的解释起步未久,针对一般场景的解释道阻且长。仍以转换器模型为例,目前针对自然语言处理和其他少数场景的解释,局限于几乎只对人工智能专家有意义的、近似模型关键权重的可视化(Abnar & Zuidema,2020;Chefer,Gur & Wolf,2021),其他人几乎不可能循此形成对模型有意义的理解。在一般场景中也有从数据表征入手解释模型的尝试,但尚未超越纯粹理论层面(Guan,Wang & Zhang et al. ,2019)。① 既然数据表征难以解释、无法理解,那么人对原始数据和模型中间环节的理解便随之断裂,人也就在很大程度上被推到了环外。挑战随之而来。就人工智能治理中公认的价值而言(Jobin,Ienca & Vayena,2019),数据表征的可理解性缺失直接导致透明性、可解释性和负责任性等价值难以实现,也会造成隐私保护、安全性和公平性等价值的实现困难。

数据表征对这些价值造成的挑战,可以在对应的法律视野下进一步展开。当前而言,虽然存在专门针对人工智能的立法提案以及其他相关适用法律,但最为全面、系统、现实地体现这些价值的法律仍然是关于个人信息保护的法律。② 在国内主要是《个人信息保护法》,在国外主要是《通用数据保护条例》(GDPR)和其他类似法律。如果这些法律适用于数据表征,那么数据表征的处理者既需要在具体处理活动中遵从透明性等原则,也需要遵从一系列体现相应原则的义务,包括体现透明性的告知义务、体现可解释性的算法说明义务、体现负责任性的自证合规义务,以及体现隐私保护、安全性和公平性的其他义务,等等。③ 据此,在现行法律下,数据表征是否属于个人信息,以及相应地是否处在个人信息保护法律的适用范围内,是其法律治理的核心问题。或者说,如果特定的数据表征属于个人信息,则其处理者需要满足这些义务,特别是需要克服数据表征缺乏可解释性的问题,提供告知、说明并自证其对数据表征的处理符合规定等。既然数据表征及其所促进的人工智能技术发展带来亟须应对的治理挑战,个

① 简言之,相应尝试的思路是将模型分解成若干表征及其相互作用(Guan,Wang & Zhang et al. ,2019)。

② 国内外未来如果通过关于人工智能治理的专门立法,则有可能改变此处论断。不过,目前而言,除了仍处于立法进程、通过尚待时日的欧盟《人工智能法案》,此类专门立法还很难与个人信息保护法律比肩。

③ 此述义务均可在《个人信息保护法》、GDPR 与其他个人信息保护法律中找到。

人信息保护法律又是目前最为全面、系统、现实的工具，无怪乎国内外立法者、执法者都在探索通过治理数据表征来加强人工智能治理。与此同时，鉴于这些法律已经适用于原始数据，但法律仍然很少涉及技术复杂的模型，通过治理数据表征还可以弥补其间的断裂。国内外的这些探索及其教益，正是即将展开的主要内容。

在此之前有两项剩余的说明。一是在数据表征产生显著变化的过程中，出现了许多含义与其相近的术语。无论是人工智能技术视角下的讨论，还是法律视角下的讨论，包括国内外立法与执法，都未能就术语使用取得统一。为避免遗漏和混淆，需要列举这些术语。上文提到了“特征”“标签”“兴趣点”和“矢量”，本节提到了“表示”，这些都是含义相近的术语。[①] 国外执法中出现的相近术语还包括“模板”[②]和“嵌入”（OAIC，2021；Federal Trade Commission［FTC］，2021a）。下文除非特别说明，否则将不加区分地使用这些术语。二是数据表征与作为人工智能前沿趋势的大模型和开源模型社区存在密切联系。受到广泛认可的大模型为数据表征提供了标准化的实现。开源模型社区在通过开源进一步促进相应实现方案传播的同时，也在探索自律治理的制度安排，包括在授权时要求用户承诺负责任地使用表征和模型（Hugging Face，2022a；Hugging Face，2022b）。在将数据表征法律治理置于更加多元的治理方式中考察并展望其未来发展时，这是有用的事实。

二、国内数据表征相关立法及其适用

国内旨在治理数据表征的法律——主要是个人信息保护的法律——分散在不同位阶中。位阶最高的《个人信息保护法》和其他相关法律法规尚未明确涵盖数据表征，其解释与适用也缺乏对于数据表征的明确立场。位阶较低的《管理规定》明确涉及数据表征，但其具体适用仍处于探索进程中，除算法备案以外暂时缺乏实例。位阶更低的技术标准[③]，特别是针对人脸识别、步态识别等生物识别技术的技术标准亦明确涉及数据表征。技术标准的实践影响相对突出，但

① 这些术语彼此之间存在着一定程度的差异，特别是使用“特征”或“标签”这两个术语时，所指的常常是可理解性较强的表征。

② 这一术语通常仅用于生物识别，特别是人脸识别的场景。

③ 国内个人信息保护领域的技术标准本身通常并不具备强制力。理论上来说，其更多的是在成本考量之下“对事实的重构或建构”（陈伟，2022）。不过，即使技术标准本身不具备强制力，实践中其仍然可能通过处理者自愿遵守或受到其他强制性规范援引的方式产生强制力（许可，2019）。《个人信息安全规范》等少数特别重要的技术标准也确实因此具备相当的影响。

仍然受制于上位法律的不明确。

《个人信息保护法》中的“个人信息”尚未明确涵盖数据表征,能否解释为一般而言涵盖数据表征也不明确,只能针对具体场景做出推测。按其中第四条,个人信息指“与已识别或者可识别的自然人有关的各种信息”。权威法律释义还从肯定和否定两方面展开了这一定义:一是为了判断哪些信息属于个人信息,需要考虑“识别”和“关联”,无论是能够通过相应信息识别个体身份,还是相应信息与处理者已经识别其身份的个体有关,又或者是相应信息与处理者能够识别其身份的个体有关(杨合庆,2022);二是为了判断哪些信息不属于个人信息①,需要考虑其是否需要付出高额成本才能复原(杨合庆,2022)。对于人脸识别等少数场景,因其所生成表征的处理目的便是识别个体,实践中也具备精准的识别能力,似可推测其属于个人信息(赵精武,2020;赵精武,2022)。尽管如此,由于相应表征的识别能力需要结合特定的硬件和模型才能实现,处理者以外的个体或组织难以通过其实现识别,针对人脸表征的上述结论仍有挑战空间(赵精武,2020;赵精武,2022)。② 对于更加一般的场景中的数据表征,例如相当常见的、推荐算法中基于用户特征的嵌入,亦即神经网络模型中特定层的输出(Goodfellow,Bengio & Courville,2016),无论是基于“识别”“关联”判断,还是基于“复原成本”分析,都很难得到确定的结论。一方面,尽管通过嵌入一般无法实现识别,但可以认为其与相应用户存在一定关联。然而,自动化生成的嵌入通常不具备直观含义,其与用户的关联因此也很难定义和测度。另一方面,嵌入通常由模型深入加工产生,一般不容易由其复原作为原始数据的用户信息,但也可能存在发生导致识别或推断的攻击的可能性(Duddu,Boutet & Shejwalker,2020)。有鉴于此,由于不同展开下都难以得到确定的解释,法律层面是否涵盖数据表征难以称为明确,只在少数场景中存在确定性稍高的结论。

作为部门规章的《管理规定》明确涉及数据表征,但其具体适用有待观察。《管理规定》的相关内容主要有四类:首先是第九条要求“加强信息安全管理,建

① 或者说属于《个人信息保护法》第七十三条第四款所定义的已经匿名化的信息。

② 人脸信息亦可通过《中华人民共和国民法典》和《最高人民法院关于审理使用人脸识别技术处理个人信息相关民事案件适用法律若干问题的规定》的进路施行理论分析和实践保护。然而,其只是简单地包含“人脸信息”这一表述,而缺乏更细致的展开,因而同样无法免疫于与此处类似的批评(赵精武,2022)。

立健全用于识别违法和不良信息的特征库，完善入库标准、规则和程序”，此处涉及特征（库）[①]；其次是第十条要求“加强用户模型和用户标签管理，完善记入用户模型的兴趣点规则和用户标签管理规则，不得将违法和不良信息关键词记入用户兴趣点或者作为用户标签并据以推送信息”，此处涉及标签和兴趣点；再次是第十七条要求“向用户提供选择或者删除用于算法推荐服务的针对其个人特征的用户标签的功能”，此处同样涉及特征和标签；最后是第二十四条建立的算法备案制度在实践遵从中可能需要描述数据表征。无论着眼国内还是国外，《管理规定》都是深度结合法律与技术的突出实例。然而目前而言，至少有三点原因制约了《管理规定》实践适用于数据表征的效果。第一是如前所述，作为《管理规定》主要上位法依据的《个人信息保护法》对于各类数据表征的解释与适用暂不明确，第十条和第十七条的相关义务因而面临是否有据、范围如何的困境。第二是即使假定立法有据，其中引入“特征（库）”“兴趣点”“标签”等术语的创新依然可谓是“双刃剑”，既是对人工智能技术实践的准确描述，又是在国内法律体系中可谓是前所未见的概念。除非能够厘清这些概念确实属于个人信息或者属于其他上位法依据中的有关范畴，并且通过解释进一步纳入“矢量”“模板”“嵌入”等同属数据表征的概念[②]，否则第九、十、十七条的适用范围都很难称为完全明确，例如第十七条是否适用于推荐算法中基于用户模型的嵌入即有争议空间。第三是少数具体条款不够清晰。仍以第十七条为例，无论此处标签是否涵盖嵌入，选择标签的含义都有待进一步澄清。是只能选择特定标签不用于后续推荐服务，还是可以选择特定标签用于后续推荐服务，抑或可以选择更精细的处理方式，其在实践上都存在实质差异。因此，虽然《管理规定》创新性地涵盖了数据表征，但为使其真正发挥效能，依然存在显著的改进空间。[③]

---

① 《互联网信息服务深度合成管理规定》第十条同样明确涉及特征（库），但仍不易推测其具体如何充分实施。

② 《管理规定》未明确其第十条和第十七条调整对象与“用户画像”的关系，是又一可能造成困惑之处。

③ 算法备案制度的顺利推进可能在一定程度上缓解这些问题。不过，尽管来自算法备案的信息足以为立法者和执法者提供背景知识和问责点，从而改进立法和执法（张凌寒，2021；许可、刘畅，2022），但仅凭这些信息不易完全修复上位法依据和概念体系等层面的问题。实际而言，仅从数据表征的角度出发，当前公示的备案信息仍然存在“标签偏好”“向量表征”“特征”“兴趣特征表征向量”等彼此各异，亦未必与《管理规定》一致的术语（北京网易传媒有限公司，2022；北京快手科技有限公司，2022），为规定适用造成了额外的困难。由此，仅就数据表征的法律治理而言，算法备案难以缓解当前立法中存在的问题。

针对具体场景中的人工智能技术的技术标准同样明确涉及数据表征，但其实践效能受到场景和位阶的双重制约。以具备代表性的《信息安全技术 人脸识别数据安全要求》（以下简称《安全要求》）为例，相比其他涉及数据表征的法律和部门规章，《安全要求》有三方面的改善：首先是第 3 条第 2 项更加明确地定义“人脸特征”为“从人脸图像提取的反映自然人脸部特征的参数”。其次是第 8 条从隐私保护和安全等价值出发，对人脸特征信息本身提出了具备可更新、不可逆和不可链接特性的要求。由此，人脸特征泄露造成的负面影响相对可控，由人脸特征出发重建人脸图像，或由其出发识别、推断个体身份相对困难。最后是第 6 条和第 8 条从隐私保护和安全等价值出发，对人脸特征的处理也提出了具体要求，包括用于生成人脸特征的人脸图像原始数据满足最小必要原则，以及生成人脸特征后立即删除相应图像。其他生物识别领域的技术标准也包含类似的、较为明确的要求。比较之下，技术标准确实更加适宜于实践。但是技术标准面临的局限同样很明显：一方面，技术标准的适用范围通常较为狭窄；另一方面，由于技术标准施加的、针对数据表征的要求本身并不具备强制力，一般只是对《个人信息保护法》或其他相关法律法规有参照意义的细化，在这些法律法规仍不明确的前提下，技术标准的定义和规范缺乏足够牢固的基础，实践效能也因此受限。为了夯实《安全要求》的基础，需要通过说理阐明《个人信息保护法》确实涵盖人脸的表征；为了克服更多技术标准面临的场景和位阶局限，需要对更加一般的数据表征阐明这一点。

综上，国内对于数据表征的立法已颇为丰富且深入，但在具体问题上，特别是个人信息是否涵盖数据表征这一问题上需要进一步夯实分析的基础。[①] 他山之石，可攻此玉。

① 国内各地法院已经审理了一些相对复杂的、涉及个人信息定性问题的案件，但这些案件通常没有脱离前述“识别”和“关联”的分析进路。以较为典型的、在《中华人民共和国民法典》相近条款下审理的“黄某与腾讯科技隐私权、个人信息权益网络侵权责任纠纷”一案（〔2019〕京 0491 民初 16142 号，通常称“微信读书案”）为例，除了基于“识别”和“关联”的分析以外，法院还指出“识别”的分析需要考虑具体场景。这样的裁判思路对于分析数据表征是有用的起点，但更进一步的结论通常需要较此更为精细的分析，包括在数据表征相关问题中细化场景这一概念。

三、国外数据表征相关执法及其说理

国外对于数据表征的法律治理主要寓于具体执法中。这些执法可以大致分为两类:一类对于数据表征是否属于、为何属于个人信息有较为详细的说理,一旦阐明这一结论,个人信息保护的法律治理循之适用于数据表征和模型;另一类则以美国 FTC 的“算法罚没”执法(Slaughter, Kopec & Batal, 2020; Li, Forthcoming)[①]为代表,虽然没有太多讨论数据表征的法律性质,但在执法中将其作为个人信息保护执法救济的一部分。

GDPR 的个人信息[②]定义及其展开,为域外涉及数据表征执法的说理提供了肯定和否定的两方面基础。首先是判断何者属于个人信息的肯定性基础。按照 GDPR 第 4 条第 1 款和“鉴于”部分第 26 条,个人信息可以定义为“与已识别或可识别自然人有关的信息”,此与《个人信息保护法》完全一致。在诺瓦克诉爱尔兰数据保护专员(Data Protection Commission[DPC],2021)一案中,欧盟法院细化了个人信息的判断依据:不仅需要抽象地考虑“识别”和“关联”,还要从“内容、目的或影响”的角度具体地分析“关联”(Court of Justice of the European Union,2017)。如此展开“识别”和“关联”的思路足以在很多执法案例中阐明数据表征为何属于个人信息。以人脸识别中的人脸矢量为例。[③] 英国和澳大利亚数据当局发现,从内容出发,矢量是人脸图像信息的“数学表征”,从目的出发,矢量的目的即进行生物识别,综合这两点足以认定人脸矢量属于个人信息(OAIC,2021)。[④] 又以广告推荐算法中表征个体隐私偏好的“透明和隐私串”(transparency & control string,以下简称 TCF)为例。比利时数据当局发现,尽管难以确定 TCF 能否识别个体,但因为处理 TCF 的目的即根据不同个体的隐

① 斯劳特尔(R. Slaughter)等人在其首先阐明“算法罚没”概念一文中的署名单位都是 FTC。

② GDPR 的表述是“个人数据”,但就本文论述目的而言没有必要区分“个人信息”与“个人数据”。

③ 对于生物识别中的数据表征(点表征)而言,存在从“识别”进路出发,仅根据其识别准确率高,且其识别能力随时间推移保持稳定即认定相应表征属于个人信息的案例(Agencia Española de Protección de Datos,2021)。不过,这一判断受到场景的局限,不易适用于生物识别以外的场景。此外还存在处理者将数据表征(模板)与其他个体标识符关联的情况(Nemzeti Adatvédelmi és Információszabadság Hatóság, 2022)。这是容易判断的情况,但同样不具备一般性。

④ 本案主要适用澳大利亚有关法律。其中的主要概念,特别是“个人信息”的概念定义和之前讨论的个人信息保护法律都是相近的。

私偏好对其区分对待——或者说处理 TCF 的目的就蕴含了对个体的区分与识别[①],故而应当推定 TCF 属于个人信息(Autorité de Protection des Données [APD],2021)。综之,无论是人脸识别还是推荐算法,此述展开都可以提供相对明确、可以实践的说理。对于更加一般的数据表征,分析其内容、目的与影响亦有意义。比起抽象的"关联",无论是数据表征所表征的信息内容,还是其及相应模型服务的处理目的,抑或其及相应模型处理对个体的影响,都可以更加具体分析。不妨以推荐算法的模型嵌入为例,虽然问题不能一劳永逸地解决,但现在可以通过其加工产生的信息、其处理目的及其处理影响给出更明确的结论。

GDPR 还为判断何者不属于个人信息的说理提供了基础。简言之,为判断特定信息不属于个人信息,需要排除"区分""关联"与"推断"三类风险,亦即需要确定处理者无法通过相应信息将特定个体从群体中区分出来,无法在属于特定个体的不同信息间建立关联,也无法推断特定个体的有关信息(Article 29 Working Party,2014)。这一思路已经针对少数非常复杂的数据表征体现出了"威力",例如通过有损哈希算法表征个体的电话号码。虽然直接分析"识别"或"关联"都不容易,但是由于算法表征以后个体依然可以区分,以及算法表征以后依然可以推断个体的社交关系网络,爱尔兰和欧盟数据当局依然认定其属于个人信息(DPC,2021)。[②] 对于经过不止一种算法处理的信息以及基于新兴合成技术生成的信息,通过三类风险同样可以给出细致的分析(Autoriteit Persoonsgegevens,2021;Stadler,Oprisanu & Troncoso,2022)。不过,尽管这一思路可以用于分析非常复杂的信息,但其本身对法律和技术的理解有较高的要求,实践应用相对肯定性的思路更为困难。综之,这一思路适宜与"内容、目的或影响"相互补强,在大部分涉及数据表征的案件中辅助分析,在剩余特别复杂的案件中发挥更重要的作用。

美国 FTC 在新近个人信息保护执法中引入的"算法罚没"措施同样涵盖了数据表征。这一概念本身不涉及数据表征是否属于个人信息的判断,但其使得个人信息保护的执法可以涵盖数据表征。所谓"算法罚没"指的是如果作为算

① 区分(对待)属于识别的一种(APD,2021;Purtova,2022)。

② 此处表征系用于联系人匹配,亦即出于区分和识别而处理。这可以和基于三类风险的分析相互补强。

法输入的原始数据违反了个人信息保护的规定，侵犯了个人对其信息的权益，那么后续救济不仅需要删除原始数据，还需要删除由算法衍生的表征和模型（Slaughter，Kopec & Batal，2020）。这一救济措施的初衷是有效阻遏使用算法侵害个人信息权益的行为（Slaughter，Kopec & Batal，2020），客观上也使得数据表征处在个人信息保护法律的射程之内。据此，FTC 在针对相册管理应用滥用人脸图像的案件中要求应用开发者删除所有衍生自图像的人脸嵌入，包括“全部或部分地由图像中人脸提取的所有数值向量等数据”（FTC，2021b）。在回复有关这一案件的质疑时，FTC 进一步明确人脸嵌入包括“概率性”——全部或部分地由图像中人脸提取的“概率分布估计”（FTC，2021c）。因此，无论数据表征是否属于个人信息，出于在执法中提供充分救济的需要，FTC 都可以灵活地设计涉及数据表征的救济措施。[①] 相比 GDPR，这是更加具有实用导向的思路。

以上，可见域外执法针对数据表征的法律治理引入了三方面的不同思路。相比国内，深度结合个人信息法律与数据表征技术的立法在国外整体而言要更加少见；不过，通过充分地展开个人信息保护法律的“个人信息”概念，以及采取其他实用导向的救济措施，国外相比国内积累了更多涉及数据表征的执法案例。于是结合二者成为自然的思路。

四、令国内法律清晰适用于数据表征

“自上而下”地结合国内立法与国外执法经验，可以使得国内法律清晰适用于数据表征。具体来说，首先从《个人信息保护法》的解释与适用入手，通过既有司法案例和相关技术标准，可以自然地引入“内容、目的或影响”和“区分”“关联”与“推断”这两类分析思路。如前所述，由此足以对人脸识别、推荐算法和其他数据表征得出法律适用的明确结论。然后，法律层面的明确适用可以向下传导至《管理规定》等部门规章和《安全要求》等技术标准。既可夯实其基础，亦可解决前面提及的其他问题。

与《个人信息保护法》中“个人信息”有关的典型司法案例引入了“场景

① 前提当然是 FTC 本身有相当宽泛的、设计与违法行为合理相关的救济措施的权力（Slaughter，Kopec & Batal，2020）。

说”,此处的“场景”可以涵盖“内容、目的或影响”。以国内影响较大的“微信读书案”为例,其中着重考虑了相应信息是否“在特定场景下”可以还原到个体的身份信息。又以影响同样较大的“余某与北京酷车易美网络科技公司隐私权、个人信息保护纠纷案”(〔2021〕粤0192民初928号,通常称“查博士案”)为例,其中同样提及“在车辆交易场景下”对涉案信息是否属于个人信息展开分析。两案均未进一步说明何谓“场景”,但其间的场景分析实际已涵盖了“内容、目的或影响”。“微信读书案”实际上已经考虑了内容和目的:一方面,在分析读书信息是否属于个人信息时,法院考虑了读书信息“包括读书时长、最近阅读、书架、推荐书籍、读书想法等,能够反映阅读习惯、偏好等”的事实;另一方面,在分析腾讯内部生成的OPEN_ID是否属于个人信息时,法院也指出了OPEN_ID本就是用于“识别用户的识别码”。“查博士案”中法院也考虑了“内容、目的和影响”。在分析历史车况信息是否属于个人信息时,综合考虑了相应信息的内容是否涉及具体个人,以及相应信息是否用于评价具体个人的行为和状态。由此,通过场景分析引入“内容、目的和影响”不仅不会与国内司法实践冲突,还可以使当前仍然较为模糊的场景分析体系化和具体化。对于人工智能和数据表征治理而言,如此不仅能够进一步明确《个人信息保护法》中的“个人信息”概念以及这一法律的适用范围,还可以对于一般的数据表征得出类似于“推定”的结论。简言之,如果数据表征系由个人信息加工或提取,并且相应的模型是用于识别个人的目的,也会对于个人权益造成影响,则一般应当认为相应数据表征属于个人信息。对于人脸识别或者一般意义的生物识别,通过涵盖“内容、目的和影响”的场景分析足以认定其一般属于个人信息;对于推荐算法嵌入等更加复杂的数据表征,结合这一分析与算法备案中公示的“算法运行机制”和“算法目的意图”(北京网易传媒有限公司,2022;北京快手科技有限公司,2022),一般也足以给出更具体的分析和更明确的结论。

还可以通过援引技术标准细化《个人信息保护法》中关于“个人信息”的解释,从而引入“区分”“关联”与“推断”的分析思路。具言之,《信息安全技术 个人信息安全影响评估指南》(以下简称《评估指南》)附录A.4已经引入了基于这三类风险的判断思路。理论上来说,法院可以通过援引《评估指南》作为判断个人信息范围的说理依据(许可,2019);实践上来说,在“微信读书案”等案件

中,法院也已援引《信息安全技术 个人信息安全规范》作为判决依据之一。[①] 于是,法院通过援引《评估指南》引入“区分”“关联”与“推断”思路,从而在涉及数据标准的案件,特别是涉及复杂数据表征的案件中给出更加具体的分析,理论上可行,实践中亦可期待。如前所述,这也可以与体系化、具体化以后的场景分析相互补充。特别是对于认为基于特定个人信息提取的数据表征不属于个人信息的主张,可以通过全面考虑三类识别风险予以审慎立场的分析。

明确个人信息一般属于个人信息且为何一般属于个人信息以后,《管理条例》和技术标准所面临的问题都可以在三个层面上得到相当程度的纾解。首先,无论是《管理条例》和技术标准中出现的“特征(库)”“兴趣点”“标签”等术语,还是法律和技术中可能出现的“矢量”“模板”“嵌入”等更多术语,都可以通过上述分析认定其一般属于个人信息,而不至于导致困惑。当然,在不同位阶立法中统一描述数据表征的术语,同样有促进明确的效果。其次,明确这些术语都属于个人信息以后,《管理条例》、技术标准和其他相关文件所设定的相关规范都具有了充分的依据。在夯实基础的同时,也为其适用范围提供了更多的确定性——以《个人信息保护法》作为上位法依据的规范的适用对象是属于个人信息的数据表征,也主要限于这些数据标准。最后,夯实依据以后,法律法规和技术标准的起草者可以更加放心地设定更多数据表征的有关规范,法院也可以更加放心地援引这些规范作为裁判或说理依据。《个人信息保护法》和其他相关法律法规体现的透明性、可解释性、负责任性、隐私保护、安全性和公平性等人工智能治理价值,因而可能得到更充分地展开和贯彻。此外,部门规章和技术标准也可以“反哺”《个人信息保护法》的解释和适用。如前,算法备案制度要求公示的信息为体系化、具体化以后的场景分析提供了有用的事实;部门规章和技术标准在坚实基础上的不断细化和完善,为解决具体争议提供了更多的裁判和说理依据。

综上,国内既有立法和国外新近执法可以自然地结合,从而为人脸识别、推荐算法以及更加一般的人工智能技术所涉及的数据表征法律治理提出相对完善的解决方案。个人信息保护体现的人工智能治理价值因而适用于数据表征,这

① 王新锐和罗为(2020)也从律师的角度指出:“在我们接触的与个人信息相关的诉讼案件中,原被告双方通常都会就是否违反《规范》进行辩论。”《规范》即《信息安全技术 个人信息安全规范》。

是仅针对现行法律的结论。通过讨论更加多元的、治理数据表征的方式,可以进一步丰富解决方案,并且增强数据表征治理作为人工智能治理“桥梁”的作用。

五、通过数据表征治理促进人工智能治理

更加多元的治理方式不仅包括法律,还包括规范、市场和架构,分别对应于社群自律、市场竞争和技术治理(Lessig,2006)。对数据表征而言,开源模型社区和大模型共同构成人工智能技术发展的前沿趋势,同时也体现了规范、市场和架构治理的重合。未来更加丰富的数据表征法律治理,需要包括与开源模型社区的充分协同。在专门立法中体现这一协同不仅有助于促进数据表征的法律治理,还有助于促进人工智能技术的法律治理。

大模型,包括前文提及的 GPT-2 和 GPT-3 等模型,既与数据表征学习的进步有关,本身又是当前人工智能治理的核心问题(Bommasani,Hudson & Adeli et al. ,2021)。由于大模型在通用任务上体现出很强的性能,具备实用价值,同时其对于计算资源和工程能力有较高的要求,而个体和大部分组织难以具备相应条件,因此以 Hugging Face 为代表的开源模型社区宣称致力于“为每个人推动和民主化”这些大模型,特别是通过开源表征预训练模型、压缩后模型和大模型接口的方式,使得个体和大部分组织可以自主学习、训练和应用大模型(Hugging Face,2022c)。开源模型社区同时通过社区自律的方式治理表征和模型的使用。例如在应用大模型之前承诺将保护个人信息,避免通过自动化决策损害他人的法律权益,避免散布虚假的或未标注其系合成的合成信息,避免损害弱势群体的权益,等等(Hugging Face,2022b)。又如在开源大模型时配备数据卡和模型卡,其中数据卡着重介绍原始数据的来源及其合法合规情况,也涵盖包括提取数据表征在内的后续处理步骤(Gebru,Morgenstein & Vecchione et al. ,2021;Hugging Face,2022d);模型卡着重介绍模型的训练、应用和评估方式,也涵盖训练过程涉及的表征学习步骤(Mitchell,Simone & Zaldivar et al. ,2019;Hugging Face,2022e)。一定程度上,甚至可以认为开源模型社区的数据卡和模型卡构成了技术色彩更鲜明,但内容也更加全面的自律算法备案。此外,开源模型社区的发展既是大模型领域市场竞争结果的体现,又使得治理风险和治理责任向有能力开

发、训练和开源维护大模型的少数组织集中。这些都是未来法律治理所面临的趋势。

站在超越现行法律、聚焦新兴领域立法的角度，综合所有前文，相应可以展望未来人工智能治理立法如何将数据表征作为治理关键环节的五个要点。第一点是进一步明确提取自个人信息的数据表征一般属于个人信息，同时确认体现人工智能治理价值的各类个人信息保护责任和义务均适用于相应数据表征。例如《个人信息保护法》第二十四条对利用个人信息决策时“保证决策的透明度和结果公平、公正”的要求，即可以数据表征作为连接点进一步适用于人工智能技术。第二点是在算法备案制度实践的基础上，吸收数据卡和模型卡的经验，进一步加强关于人工智能技术的信息披露。特别是对于数据表征而言，需要描述其提取、加工和学习的表征类型以及相应机制，并且简要描述表征的内容。[①] 这一点可以作为备案中“算法运行机制”的一部分，和备案所包括的“算法目的意图”，以及“（尤其是对个人权益影响的）算法自评估报告”一同支持执法和司法中的法律适用。第三点是认可和鼓励基于社区自律的规范治理：一方面，在立法中明确有关数据表征和相关模型开发、训练与应用的自行承诺对承诺者具有约束力，可以通过行政执法和公益诉讼等方式执行；另一方面，通过尽职免责、认可作为社会责任、认可作为责任承担的减轻因素等方式加强自律的正面激励。第四点是将“算法罚没”，包括删除数据表征和相关模型，作为责任承担方式之一。如前所述，从实用导向的角度看，这一责任具有可观的阻遏效果。第五点是考虑到以大模型为代表的人工智能技术权力和治理风险愈发向以“守门人”为主的少数组织集中，一方面是需要在人工智能治理立法中引入关于守门人主体的增强规定，另一方面是如果通过针对守门人的专门立法，例如平台法中也应当包括人工智能治理的有关内容。在更加清晰适用现行法律的基础上，上述五点和规范、市场与技术共同构成了数据表征和人工智能治理更完善的方案。

① 这一描述可以是相对技术性的，例如特定的数据表征提取了相隔较远的字词间的语义联系。

六、结语

以上首先概述数据表征的发展为何对人工智能技术进步有重要作用,以及人工智能治理如何尝试将数据表征作为原始数据和下游模型间的桥梁。然后,一方面,评述了国内正在日趋完善的、涉及数据表征的立法;另一方面,评述了国外日益丰富的、令法律适用于数据表征的执法实践。之后结合二者,“引石攻玉”,自然将域外实践发展的有用思路引入国内法律,使得国内法律可能清晰适用于数据表征,从而贯彻人工智能治理追求的诸多价值。最后,对于立法的展望虽然只是在技术前沿下的初步推测,但我们也坚信其将很快成为现实。

参考文献

北京快手科技有限公司,2022,《网信算备 11010841376070222001 7 号》,https://beian.cac.gov.cn/api/static/fileUpload/principalOrithm/additional/user_bb4fec44-be41-43d7-a11e-6fcd8d7776ad_9ea7387d-c989-4625-9bb3-2e23b29a7a0c.pdf。

北京网易传媒有限公司,2022,《网信算备 11010111352243022200 29 号》,https://beian.cac.gov.cn/api/static/fileUpload/principalOrithm/additional/user_b6c4caf9-5a05-4387-bdbb-850ddb4d61cc_4f8a9fc8-6995-48da-8bee-c07cd97f8619.pdf。

陈伟,2022,《作为规范的技术标准及其与法律的关系》,《法学研究》第 5 期。

厄兹代米尔,锡南、迪夫娅·苏萨拉,2019,《特征工程:入门与实践》,庄嘉盛译,北京:人民邮电出版社。

王新锐、罗为,2020,《〈个人信息安全规范〉被援引情况实证分析》,https://www.leb-china.com/upload/upload/608dfa165de3833c47afe6ad4a194bcd.pdf。

许可,2019,《〈个人信息安全规范〉的效力与功能》,《中国信息安全》第 3 期。

许可、刘畅,2022,《理解算法备案》,《人工智能》第 1 期。

杨合庆,2022,《中华人民共和国个人信息保护法释义》,北京:法律出版社。

张凌寒,2021,《网络平台监管的算法问责制构建》,《东方法学》第 3 期。

赵精武,2020,《〈民法典〉视野下人脸识别信息的权益归属与保护路径》,《北京航空航天大学学报》(社会科学版)第 5 期。

赵精武,2022,《刷脸问题治理逻辑的一般规则转向——以技术透明度为基本立场》,《北方法学》第 1 期。

周志华,2016,《机器学习》,北京:清华大学出版社。

Abnar, S. & W. Zuidema 2020, "Quantifying Attention flow in Transformers." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Agencia Española de Protección de Datos 2021, "Expediente N°: PS/00050/2021." https://www.aepd.es/es/documento/ps-00050-2021.pdf.

Article 29 Working Party 2014, "Opinion 05/2014 on Anonymisation Techniques." https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

Australian Information Commissioner and Privacy Commissioner 2021, "Commissioner Initiated Investigation of Clearview AI, Inc. (Privacy) [2021] AICmr 54 (14 October 2021)." https://www.oaic.gov.au/__data/assets/pdf_file/0016/11284/Commissioner-initiated-investigation-into-Clearview-AI,-Inc.-Privacy-2021-AICmr-54-14-October-2021.pdf.

Autorité de Protection des Données 2021, "Decision on the Merits 21/2022 of 2 February 2022." https://www.autoriteprotectiondonnees.be/publications/decision-quant-au-fond-n-21-2022-english.pdf.

Autoriteit Persoonsgegevens 2021, "Gemeente Enschede." https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/boetebesluit_ap_gemeente_enschede.pdf.

Bengio, Y., A. Courville & P. Vincent 2013, "Representation Learning: A Review and New Perspectives." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8).

Bommasani, R., D. Hudson & E. Adeli et al. 2021, "On the Opportunities and Risks of Foundation Models." arXiv preprint arXiv: 2108.07258.

Brown, T., B. Mann & N. Ryder et al. 2020, "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33.

Chefer, H., S. Gur & L. Wolf 2021, "Transformer Interpretability Beyond Attention Visualization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Court of Justice of the European Union 2017, "Peter Nowak v Data Protection Commissioner." https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62016CJ0434.

Data Protection Commission 2021, "In the Matter of WhatsApp Ireland Limited." https://www.dataprotection.ie/sites/default/files/uploads/2022-03/Full_decision_WhatsApp_Ireland-August_2021.pdf.

Devlin, J., M. Chang & K. Lee et al. 2019, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." Proceedings of the 2019 Conference of the North American

Chapter of the Association for Computational Linguistics: Human Language Technologies.

Duddu, V., A. Boutet & V. Shejwalkar 2020, "Quantifying Privacy Leakage in Graph Embedding." MobiQuitous 2020 – 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services.

Federal Trade Commission 2021a. "California Company Settles FTC Allegations It Deceived Consumers about Use of Facial Recognition in Photo Storage App." https://www.ftc.gov/news-events/news/press-releases/2021/01/california-company-settles-ftc-allegations-it-deceived-consumers-about-use-facial-recognition-photo.

Federal Trade Commission 2021b, "Decision and Order." https://www.ftc.gov/system/files/documents/cases/1923172_-_everalbum_decision_final.pdf.

Federal Trade Commission 2021c, "Letter to Commenter World Privacy Forum." https://www.ftc.gov/system/files/documents/cases/wpf_response_final_0.pdf.

Gebru, T., J. Morgenstein & B. Vecchione et al. 2021, "Datasheets for Datasets." *Communications of the ACM* 64(12).

Goodfellow, I., Y. Bengio & A. Courville 2016, *Deep Learning*, Cambridge: MIT Press.

Guan C., X. Wang & Q. Zhang et al. 2019, "Towards A Deep and Unified Understanding of Deep Neural Models in NLP." International Conference on Machine Learning.

Hugging Face 2022a, "Getting Started With Embeddings." https://huggingface.co/blog/getting-started-with-embeddings.

Hugging Face 2022b, "Big Science RAIL License v1.0." https://huggingface.co/spaces/bigscience/license.

Hugging Face 2022c, "The AI Community Building the Future." https://huggingface.co.

Hugging Face 2022d, "Create A Dataset Card." https://huggingface.co/docs/datasets/dataset_card.

Hugging Face 2022e, "Building A Model Card." https://huggingface.co/course/chapter4/4?fw=pt.

Jobin, A., M. Ienca & E. Vayena 2019, "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1(9).

LeCun, Y., Y. Bengio & G. Hinton 2015, "Deep Learning." *Nature* 521(7553).

Lessig, L. 2006, *Code Version 2.0.* New York: Basic Books.

Li, T. Forthcoming, "Algorithmic Destruction." *Southern Methodist University Law Review*.

Liu, Y., M. Ott & N. Goyal et al. 2019, "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv preprint arXiv: 1907.11692.

Mitchell, M., W. Simone & A. Zaldivar et al. 2019, "Model Cards for Model Reporting." Proceedings of the Conference on Fairness, Accountability, and Transparency.

Nemzeti Adatvédelmi és Információszabadság Hatóság 2022, "NAIH-85-3/2022." https://naih.hu//hatarozatok-vegzesek? download = 517: mesterseges-intelligencia-alkalmazasanak-adatvedelmi-kerdesei.

Purtova, N. 2022, "From Knowing by Name to Targeting: The Meaning of Identification under the GDPR." *International Data Privacy Law* 12(3).

Radford, A., W. Jeffrey & R. Child et al. 2019, "Language Models Are Unsupervised Multitask Learners." *OpenAI Blog* 1(8).

Slaughter, R., J. Kopec & M. Batal 2020, "Algorithms and Economic Justice: A Taxonomy of Harms and A Path Forward for the Federal Trade Commission." *Yale Journal of Law & Technology* 23(1).

Solaiman, I., M. Brundage & J. Clark et al. 2019, "Release Strategies and the Social Impacts of Language Models." arXiv preprint arXiv: 1908.09203.

Stadler, T., B. Oprisanu & C. Troncoso 2022, "Synthetic Data-Anonymisation Groundhog Day." 31st USENIX Security Symposium (USENIX Security 22).

Vaswani, A., N. Shazeer & N. Parmar et al. 2017, "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30.