基于信任的自动化决策：算法解释权的原理反思与制度重构

丁晓东*

内容提要　基于个人信息的算法自动化决策越来越普遍，我国《个人信息保护法》和相关国内外法律都引入算法解释权加以应对。但由于算法具有黑箱性、不确定性与复杂性，算法解释权是否应理解为一般性权利还是限定性权利，要求系统性解释还是个案性解释、事前解释还是事后解释、人工解释还是机器解释，都存在解释争议与适用困境。在原理层面，这一争议与困境源于个人算法控制论。应以沟通信任原则重构算法解释权，将算法解释权的性质视为一种程序性权利，而非实体性权利。算法解释权制度也应进行重构，建议根据自动化决策所处的行业领域、市场化程度、个案影响、企业能力而对其内容、程度、时间和方式作不同要求。

关键词　算法　自动化决策　算法黑箱　算法解释权　个人信息

目　次

*　中国人民大学法学院、未来法治研究院副教授，法学博士。本文系 2018 年度国家社科基金一般项目“大数据背景下的个人信息保护与企业数据权属研究”（项目批准号：18BFX198）的阶段性成果。

一、引 言

大数据与人工智能时代，基于个人信息的算法自动化决策越来越普遍。如，在网约车的场景下，算法可以为乘客派送不同的司机，设置不同的线路；在资讯分发的场景下，算法可以为不同用户推荐不同的视频、新闻与广告；在搜索场景下，算法可以为每个用户呈现不同的搜索结果；在电商场景下，算法可以结合用户的搜索、浏览、购买等记录，为用户推荐符合其偏好的商品。算法自动化决策在给个人带来便捷服务的同时，其黑箱属性和可解释性也引起了社会的普遍关注。[①] 当乘客利用网约车打车时，乘客可能希望知晓，算法为何给其派送了较远的车辆，而给其他乘客派送了较近的车辆。当外卖骑手被困算法系统，不断被压缩送餐时间的外卖骑手可能希望了解，算法如何对其进行考核。[②] 当货车司机因为防止疲劳驾驶的自动记录仪掉线而被罚款，货车司机可能希望了解，此类自动化决策仪器的算法如何设置。[③] 当特斯拉等自动驾驶车辆遭遇刹车失灵，此时车主可能希望了解，车辆所遭遇的问题是车辆的问题还是自身的问题。[④]

为了应对与个人信息相关的算法黑箱问题，各国都在个人信息或个人数据保护法中规定了相关的算法解释权条款。[⑤] 如，我国《个人信息保护法》在第 24 条等条款对拒绝自动化决策的权利与解释说明的权利进行了规定[⑥]，国家互联网信息办公室制定的《互联网信息服务算法推荐管理规定（征求意见稿）》第 12 条规定，算法推荐服务提供者应保证和优化算法决策规则的透明度和可解释性；欧盟《一般数据保护条例》也在若干条款中对算法解释权进行了规定。

但是，有关算法解释权的内容、程度、时间、方式等制度问题仍然存在很大争议。就内容而言，算法解释权应视为一般性的算法解释主张，还是应视为拒绝自动化决策等限定权利主张？就程度而言，个体可否要求自动化决策者对算法进行系统性解释，还是可以要求更为具体的个案解释？就时间而言，个体是否可以要求自动化决策者在算法决策

① 参见张文显：《构建智能社会的法律秩序》，载《东方法学》2020 年第 5 期，第 4 页。

② 参见赖祐萱：《外卖骑手，困在系统里》，载《人物》2020 年第 8 期，第 70 页。

③ 参见《“北斗掉线”司机自杀背后：北斗概念正被滥用无正常数据就会罚》，载新浪网，http://finance.sina.com.cn/tech/2021-04-09/doc-ikmyaawa8632222.shtml，2021 年 6 月 20 日访问。

④ 参见《〈人民日报〉评论：尊重消费者，特斯拉该补上这一课》，载新浪网，http://finance.sina.com.cn/tech/2021-04-23/doc-ikmxzfmk8434426.shtml，2021 年 6 月 20 日访问。

⑤ 为了讨论需要，本文采用广义上的算法解释权界定。所谓广义上的算法解释权，指的是法律赋予个体的与算法决策相关的各类权利。例如拒绝纯自动化决策的权利，本文也将其视为算法解释权加以讨论。狭义的算法解释权指的是，个人有权对任何算法决策主张解释权。*See* Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, International Data Privacy Law, Vol.7:20, p.20-21(2017).

⑥ 其他法律法规也对算法解释权或脱离自动化决策权进行了一些规定，但这些规定也可以视为个人信息保护在其他法律中的体现。如，《电子商务法》第 18 条规定，电子商务经营者根据消费者的兴趣爱好、消费习惯等特征向其提供商品或者服务的搜索结果的，应当同时向该消费者提供不针对其个人特征的选项。

前进行算法解释，还是只能要求算法决策的事后解释？就方式而言，当个人提起算法解释，自动化决策者是否必须对个体提供人工解释，还是也可以提供机器解释？

目前，学术界已经对部分上述问题进行了研究。有的研究探讨《个人信息保护法》是否引入了一般性的算法解释权，⑦ 有的研究分析欧盟《一般数据保护条例》中算法解释权的性质，⑧ 有的研究从算法治理的角度进行了分析，⑨ 还有的研究特定行业的算法解释权 ⑩ 等。这些研究指出，算法解释权面临争议和挑战，并且从不同角度提出了解释方法与应用对策。本文试图对算法解释权的解释争议与应用困境作更系统性阐述，在原理层面借鉴信任与沟通机制理论对算法解释权进行新的反思，并在此基础上进行制度重构。

具体而言，本文认为算法解释权不应建立在算法个体控制的基础之上。因为算法的复杂性、不确定性以及场景多元性，强化个体算法控制面临种种挑战。作为替代，应以信任机制重构自动化决策中的算法解释权，将算法解释权视为一种信任沟通机制。当相关算法解释权主张有利于促进二者之间的信任关系时，此时应支持算法解释权的主张；相反，当算法解释权无助于二者信任关系建构时，或者当算法解释权主张挑拨二者关系，则应审慎对待算法解释权的主张。从信任沟通机制出发，算法解释权的性质应被视为一种相对性与程序性权利，而非绝对性和实体性权利。在制度层面，可以根据算法自动化决策所位于的行业领域、市场化程度、个案影响、企业能力而对算法解释权的内容、程度、时间和方式作不同要求。⑪

二、算法解释权的解释争议

目前各国都在其个人信息或个人数据保护法中引入了算法决策的相关条款，但由于与算法相关的权利相对新颖，而且各国的法律规定都比较概括抽象，这些条款往往可以作不同解读。研究者也对算法解释权的权利内容、程度、时间与方式给出了不同的法律解释。不同的算法解释权意味着个人可以主张行使的权利不同，个人信息处理者应当承担的义务不同。本部分结合各国算法解释权的立法与研究状况，分析算法解释权在法律解释层面的争议。

---

⑦ 参见吕炳斌：《论个人信息处理者的算法说明义务》，载《现代法学》2021 年第 4 期。

⑧ *See* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, Fordham Law Review, Vol.87:1085, p.1118–1126(2018); Bryan Casey, Ashkon Farhangi & Roland Vogl, *Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise*, Berkeley Technology Law Journal, Vol.34:143, p.145 (2019); Maja Brkan, *Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond*, International Journal of Law and Information Technology, Vol.27:91, p.91–93(2019).

⑨ 参见王莹：《算法侵害类型化研究与法律应对——以〈个人信息保护法〉为基点的算法规制扩展构想》，载《法制与社会发展》2021 年第 6 期；张吉豫：《构建多元共治的算法治理体系》，载《法律科学》2022 年第 1 期。

⑩ 参见张凌寒：《商业自动化决策的算法解释权研究》，载《法律科学》2018 年第 3 期。

⑪ 算法规制的场景化进路，可参见丁晓东：《论算法的法律规制》，载《中国社会科学》2020 年第 12 期，第 150–152 页。

（一）一般主张与限定主张

在权利内容方面，算法解释权一方面可以被解释为一般性权利主张，即此种权利赋予了个体在所有情形下针对所有自动化决策者的算法解释权利。以我国《个人信息保护法》为例。《个人信息保护法》第 24 条是对自动化决策进行规定的专门条款，在第 1 款中规定，“个人信息处理者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”。此外，第 7 条规定：“处理个人信息应当遵循公开、透明原则，公开个人信息处理规则，明示处理的目的、方式和范围。”第 44 条规定，“个人对其个人信息的处理享有知情权、决定权，有权限制或者拒绝他人对其个人信息进行处理”。综合这些条款，可以从中推导出一种基于透明性主张的算法解释权，个人可以对所有算法决策者提出算法解释与说明的主张，并且可以在信息处理者拒绝的情形下提起诉讼请求。另一方面，《个人信息保护法》的相关规定又可以被理解为限定性的权利性主张，包括特定条件下的说明解释权与拒绝自动化处理的权利，但不包括一般情形下的算法解释权。在这种观点看来，《个人信息保护法》第 24 条第 1 款仅一般性地规定了自动化决策的透明性要求，第 7 条和第 44 条则更是关于个人信息处理的一般性规定。而第 24 条第 2 款、第 3 款则明确对算法解释权进行了限定，这两款分别规定，“通过自动化决策方式向个人进行信息推送、商业营销，应当同时提供不针对其个人特征的选项，或者向个人提供便捷的拒绝方式”“通过自动化决策方式作出对个人权益有重大影响的决定，个人有权要求个人信息处理者予以说明，并有权拒绝个人信息处理者仅通过自动化决策的方式作出决定”。这两款将算法解释权限定为信息推送与商业营销场景下拒绝算法决策的权利，以及个人权益受到重大影响前提下得到说明与拒绝算法决策的权利。⑫

欧盟《一般数据保护条例》具有一定的相似性。一方面，《一般数据保护条例》在第 3 章第 2 部分“信息与对个人数据的访问”的多个条文中对算法透明进行了规定：当数据控制者“收集数据主体个人数据时”（第 13 条），“当个人数据还没有从数据主体那里收集”（第 14 条）以及当数据主体的“数据正在被处理时”（第 15 条），数据主体有权知晓“存在自动化的决策”，包括“用户画像和在此类情形下，对于相关逻辑包括此类处理对于数据主体的预期后果的有效信息”。此外，《一般数据保护条例》还规定，对于所有个人数据权利，“控制者应当以一种简洁、透明、易懂和容易获取的形式，以清晰和平白的语言来提供”（第 12 条）。有关研究认为，从这些条款中可以推导出一种一般性的算法解释权，因为离开了算法解释权，用户的知情权就不能保证，也不可能获取“对于数据主体的预期后果的有效信息”。⑬ 另一方面，欧盟《一般数据保护条例》中的算法解释权也可以被视为一种限定性权利。按照这种观点，《一般数据保护条例》第 12—15 条的相关规定只是对透明性的要求，而第 21 条与第 22 条对算法的专门性规定只赋予了

⑫ 参见贾章范：《论算法解释权不是一项法律权利——兼评〈个人信息保护法（草案）〉第二十五条》，载《电子知识产权》2020 年第 12 期，第 60–61 页。

⑬ *See* Gianclaudio Malgieri & Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, International Data Privacy Law, Vol.7:243, p.243–246(2017).

个体有限的权利。其中第21条规定了数据主体在用户画像与营销中使用算法的反对权，即“因为直接营销目的而处理个人数据，数据主体有权随时反对为了此类营销而处理相关个人数据，包括反对和此类直接营销相关的用户画像。数据主体反对为了直接营销目的而处理，将不能为了此类目的而处理个人数据”。第22条规定了限定情形下对基于自动化的个人决策和用户画像的反对权，数据主体有权反对“完全依靠自动化处理——包括用户画像——对数据主体作出具有法律影响或类似严重影响的决策”。

算法解释权的一般性主张与限定性主张之间的区别并不仅仅是概念性的，前者比后者为个人赋予了更宽泛的权利主张，其适用对象更广，而且对算法决策者施加了更多责任。如，在一般娱乐资讯的个性化推荐机制下，按前者理解，个体可以要求企业对其自动化决策算法进行解释，但按后者理解，则个体仅具有拒绝自动化决策或关闭个性化推荐的权利，对于没有给个人造成严重影响的算法决策，个人无权主张一般性的算法解释权。

（二）系统解释与个案解释

算法解释权的程度要求也可以作两种理解。一种理解是，个人可以要求自动化决策者对算法的系统功能进行解释，即要求自动化决策者解释自动决策系统的逻辑、意义、预期后果和一般功能，如系统的需求规范、决策树、预定义模型、标准和分类结构。另一种理解是，个人可以要求对个案决策进行解释，即要求自动化决策者解释“算法决策的基本原理、原因，以及限定自动决策的个别情况，如特征权重、机器定义的案例限定决策规则、有关引用或配置文件组的信息”⑭。两种解释为个人赋予不同的信息权利，为自动化决策者施加不同的责任。以网约车场景为例，按照前一种理解，乘客有权知晓网约车公司利用算法进行派车的一般机制，网约车公司需要为个体提供其算法决策规则的一般性解释。而按照后一种理解，乘客则可以要求网约车就某次车辆派送进行个案性解释，网约车公司不但需要解释其算法决策的一般性规则，而且需要向某位乘客解释，为何给其派送某种类型的车辆，为何其等待的时间较长，为何给其设置了某条行车路线。⑮

从法条解释来看，以上两种程度要求都可以在我国的《个人信息保护法》中找到依据。一方面，《个人信息保护法》中的算法解释权可以被视为一种系统性解释要求。《个人信息保护法》第48条规定，个人有权要求“对其个人信息处理规则进行解释说明”。这意味着，自动化决策者仅仅需要对其算法规则进行解释，而不需要对具体的决策机制进行说明。另一方面，我国《个人信息保护法》中的算法解释权可以被视为一种个案性解释。《个人信息保护法》第24条规定，“应当保证决策的透明度和结果公平、公正”“通过自动化决策方式作出对个人权益有重大影响的决定，个人有权要求个人信息处理者予以说明”，这一规定似乎不仅仅针对一般性的规则，而且还包括对个案的算法决策进行说明。

欧盟的《一般数据保护条例》也可以作两种解释。一方面，《一般数据保护条例》可以被视为赋予了个体以系统解释权：第13、14、15条赋予了个人对于“存在自动化决

⑭ Sandra Wachter, *supra* note 5,78.

⑮ 参见张恩典：《大数据时代的算法解释权：背景、逻辑与构造》，载《法学论坛》2019年第4期，第153-155页。

策”的一般知情权。另一方面,第 22 条第 3 款规定“数据控制者应当采取适当措施保障数据主体的权利、自由、正当利益,以及数据主体对控制者进行人工干涉,以便表达其观点和对决策进行异议的基本权利”,这一规定又明确赋予了个体以介入具体算法决策的权利。因此,《一般数据保护条例》也可以被视为赋予了个体以个案性解释的权利。

(三)事前解释与事后解释

就解释的时间要求而言,算法解释权既可以视为要求自动化决策者进行算法决策前的解释,也可以视为要求自动化决策者进行算法决策后的解释。根据事前解释的要求,无论是网约车乘客还是各类资讯、广告、新闻的用户,都可以要求自动化决策者在算法决策前就公布其算法规则。而根据事后解释,则个体只能在算法决策作出之后提出请求。

对于算法解释权的时间要求,我国《个人信息保护法》并未进行明确规定。《个人信息保护法》第 24 条、第 48 条和第 50 条中对透明性的一般性要求,既可以被解读为要求自动化决策者进行事前解释,也可以被解读为要求自动化决策者进行事后解释,或者同时包含两者。另一方面,《个人信息保护法》有的条款更偏向于事前解释或事后解释。如第 24 条的规定更偏重决策本身的公平公正与个人在事后决策中的知情权。而第 55 条规定,利用个人信息进行自动化决策,“个人信息处理者应当事前进行个人信息保护影响评估,并对处理情况进行记录”,这一条款明确了风险评估应当属于事前责任。

欧盟《一般数据保护条例》也可以作两种解释。一方面,《一般数据保护条例》中的很多条款规定了收集与处理个人信息前的告知性权利,这些条款明确赋予了个体以事前解释的权利。另一方面,《一般数据保护条例》第 22 条规定的拒绝自动化处理的权利,在一定程度上赋予了个体以事后解释的权利。对《一般数据保护条例》第 22 条进行解释的“重述”(recital)中明确提到,个体拥有“获得对评估后作出的决定的解释和对决定提出质疑的权利”。“重述”虽然没有法律效力,但在一定程度上代表了欧盟对于《一般数据保护条例》进行解读的官方立场。⑯

(四)人工解释与机器解释

就解释方法而言,算法解释权可以被认为要求进行人工解释,即当个人提出算法解释权的主张时,自动化决策者应当通过人工服务的方式解释算法的运行机制。另一方面,也可以认为纯粹机器解释同样符合算法解释权的要求。在当前很多互联网公司提供的客户咨询中,都采取了机器解释的方式,即通过自动化或非人工的方式为顾客解答疑虑。

目前各国的个人信息保护法并未对解释方法作出明确规定。一方面,相关法条似乎蕴含了人工解释的要求。如,我国《个人信息保护法》第 24 条规定的个人有权拒绝“仅通过自动化决策的方式作出决定”和“有权要求个人信息处理者予以说明”以及第 48 条规定的个人有权“对其个人信息处理规则进行解释说明”,都隐含了要求进行人工解

⑯ 《一般数据保护条例》由正式法律条款与“重述”两部分组成。其中法律条款具有直接的法律效力,而“重述”则不具备直接法律效力,但为法律条款提供附加信息以及具体场景。欧洲数据保护委员会(前身为欧盟第 29 条数据保护工作组)也依赖“重述”对法律条款进行解释。此外,欧洲正义法院也经常参考“重述”,以决定《一般数据保护条例》的含义和适用。

释的要求。此外第 50 条规定的“申请受理和处理机制”，也隐含了人工介入和人工解释的要求。另一方面，相关条款也可以被理解为需要人工介入，但未必需要最终以人工方式向个人进行解释。如，银行或征信企业可以建立人工审核机制，对提出异议或拒绝算法自动化决策的用户进行人工复核，但此类复核结果却可能被录入到计算机系统，并以自动化回复的方式与用户沟通。

欧盟《一般数据保护条例》和相关规定也可以解读出人工解释与机器解释两种不同要求。一方面，《一般数据保护条例》在第 21、22 条中规定的反对用户画像与完全自动化处理的权利，在一定程度上提出了人工介入的强制性要求。如果相关沟通与解释机制是完全机器化的，则算法解释机制就可能走向其反面，变成了另一种自动化决策。但另一方面，欧盟《一般数据保护条例》并未明确说明，除了需要在设计、运行阶段对自动化决策进行人工介入之外，是否需要在与用户的沟通中也必须采取人工的方式，以及是否至少可以采用机器解释的方式。在欧盟第 29 条数据保护工作组发布的《个人自动化决策和用户画像指南》中，特别提到了使用“视频”等方式来阐述过去的自动化决策如何做出。[17] 这说明，欧盟立法者与政策制定者心目中的算法解释至少包含了机器解释。

三、算法解释权的适用困境

算法解释权的困境不仅存在于法律解释层面，也存在于法律适用与效果层面。无论将算法解释权界定为一般性权利主张还是限定性权利主张、系统性解释还是个案性解释、事前解释还是事后解释、人工解释还是机器解释，都存在法律适用的有效性难题。一旦对算法解释权的内容要求、程度要求、时间要求、方式要求作刚性的规则要求，就会带来各类问题。

（一）算法解释权的权利内容

就权利内容而言，将算法解释权界定为一般性权利主张，首先可能面临知情同意原则带来的问题。以“知情—同意”模式保护个人信息相关权益，本身就面临着很多困境，可能同时引发个人信息权益保护不足或者保护过度的问题。目前，已有不少文献对此进行了论述。如，有学者指出，对于信息处理者的告知，个人一般没有兴趣、时间、专业能力进行阅读。[18] 即使个体了解相关告知，也往往因为认知偏差而过高估计有的风险，过低判断有的风险。[19] 还有学者指出，在个人信息收集与处理中，个人信息往往涉及多个不

---

⑰ *See* Article 29 working party, *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679*, p.21, at http://ec.europa.eu/justice/data-protection/index_en.htm(Last visited on June 20, 2021).

⑱ *See* Lorrie Faith Cranor, *Necessary But Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice*, Journal on Telecommunications & High Technology Law, Vol.10:273, p.273-274(2012).

⑲ *See* Alessandro Acquisti & Jens Grossklags et al., *What Can Behavioral Economics Teach Us About Privacy?*, in Acquisti A., Vimercati S. C., Gritzalis S & Lambrinoudakis C.ed., Digital Privacy: Theory, Technologies and Practices, Auerbach Publications, Taylor and Francis Group, 2007, p.367.

同主体，其风险往往具有累积性效应（aggregating effect）。[20] 因此，个人未必能通过“知情—同意”的方式来维护自身权益，个人既可能很容易就同意信息的收集与处理，从而不能有效保护自身权益，也可能高估或者误判某些信息收集与处理行为，拒绝本来可能为个人提供更好服务或利益的相关活动。

在涉及算法的自动化决策中，上述问题更为明显。相比个人信息的收集与非自动化决策的信息处理，涉及算法的规则往往更为复杂。算法往往由很多程序员共同完成，每个程序员可能仅仅负责其中一部分，很少有人能够完整地解释算法的所有细节。尤其是在涉及大数据的算法中，由于基于大数据的算法奠基于相关关系，而非因果关系之上，对算法的解释往往难以满足人类因果关系思维的解释。[21] 在机器学习的年代，算法又增加了不确定性的特征。机器学习的特征是，运用海量数据对机器学习算法进行训练，让机器学习算法进行自我调整，以获得更好的性能。[22] 机器学习算法可以随时间的推移而改变处理数据的方式，就像人类可以通过学习改变处理数据的方式一样。因此，如果将绝对透明视为算法解释权的要求，未必现实。就效果而言，也未必能很好地维护个人的相关性权益，促使企业为个人提供更优服务。[23]

此外，将算法解释权视为一种限定性权利的主张，以拒绝算法自动化处理权作为核心，也面临不少困境与挑战。首先，在有的场景下，个人反对权的行使并不合理。如，在自动驾驶的场景下，行人主张脱离算法自动化处理，要求自动驾驶汽车避免使用算法，就不现实。因为自动驾驶在遇到行人时，必然需要即时性的算法处理。其次，在原理层面，将反对或脱离算法决策上升为一种权利也需要进一步论证。很多人可能更愿意采用自动化决策算法，而非人工算法。他们可能更加担心人工审核会带来歧视、腐败与效率低下，[24] 或者出于认知上的惰性，[25] 更愿意选择基于算法的自动化决策，而非人类决策。如果拒绝自动化决策可以作为一种权利，那么反对人类决策和采取自动化决策权是否也应该成为一种权利？

（二）算法解释权的解释程度

就算法解释权的解释程度而言，算法的系统性解释除了上文提到的算法复杂性之

---

[20] *See* Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, Harvard Law Review, Vol.126:1880, p.1889–1991(2013).

[21] 正如舍恩伯格所言，“当我们说人类是通过因果关系了解世界时，我们指的是我们在理解和解释世界各种现象时使用的两种基本方法：一种是通过快速、虚幻的因果关系，还有一种就是通过缓慢、有条不紊的因果关系。大数据会改变这两种基本方法在我们认识世界时所扮演的角色”。参见［英］维克托·迈尔·舍恩伯格、肯尼斯·库克耶：《大数据时代：生活、工作与思维的大变革》，盛杨燕、周涛译，浙江人民出版社 2013 年版，第 197–198 页。

[22] *See* Jenna Burrell, *How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms*, Big Data & Society, Vol.3:1, p.1–12(2016).

[23] *See* Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, New Media & Society, Vol.20:973, p.973(2016).

[24] *See* Anupam Chander, *The Racist Algorithm*, Michigan Law Review, Vol.115:1023, p.1044(2016).

[25] *See* Linda J. Skitka, *Does Automation Bias Decision-making*, International Journal of Human–Computer Studies, Vol.51:991, p.991(1999).

外，还存在如下几方面的难题。

首先，算法的系统性解释面临时时更新的难题。以机器学习为代表的人工智能算法可以从数据中学习，并随着时间的推移而提高其准确性。因此，在机器学习算法中，算法随着时间的演进而变化，而非一成不变。要对这样的算法进行系统性解释，就意味着必须对算法进行实时跟踪，不断进行解释。㉖

其次，算法的系统性解释常常难以引起个人注意，为个人提供有效信息。在自动化决策者进行自动化决策之前对个人的告知中，个人对算法自动化决策系统的逻辑、意义、预期后果和一般功能往往就没有多大兴趣，更不用说去了解和阅读相关告知。而在个人遭遇自动化决策，对算法产生相应疑虑之后，关于算法的一般系统解释也未必能打消个人疑虑。特别是当系统解释与个人的算法决策结果相关性不高时，系统解释对于个人就没有太多意义。

最后，算法的系统性解释还面临知识产权与“算计”算法的难题。算法的系统性解释，往往意味着算法决策机制一定程度的公开，会对企业的知识产权保护带来一定的挑战。而算法如果向包括社会公众在内的人员完全公开，则可能导致社会主体钻算法的空子，采取各种方式来“算计”算法的问题。如，Google 最初采取一种叫做 PageRank 的算法，用于在搜索引擎结果中对网页进行排名。Google 公司曾经向社会公开这种算法，但结果是有的网站利用这一算法而将自己的网站排在 Google 搜索结果前列。㉗ 为了应对这一问题，Google 只能采取其他算法与 PageRank 算法混合，不再唯一使用某种算法。㉘

相比算法的系统性解释，算法的个案性解释需要解释某个具体算法决策的逻辑，因而更贴近个人诉求。但个案性解释首先面临系统性解释同样的问题，因为个案性解释以系统性解释为前提，其需要首先解释算法决策系统的一般规则。其次，个案性解释的解释难度更大、成本更高。因为个案性解释除了要掌握算法的系统性运作，还需要掌握针对某个个体决策的具体情况。最后，算法的个案性解释需要多具体？从逻辑上说，个案性解释可以无限具体，个人可以无限追问。以个性化推荐算法为例，个人可能希望知道，为何自己收到植发广告？企业可能告诉某个个人，企业并没有设置“掉发”“秃头”之类的标签，植发广告是算法的自动化分发。但个人可能会进一步追问，企业是否设置了“头发”“中年”等标签，或者要求企业解释，为何其电脑上收到了植发广告，而手机上没有收到？为何对他推送了植发广告，而没有对其他人推送？如果个案解释以个人的彻底满意为标准，个案解释可能会无穷无尽。

㉖ *See* Fred H. Cate et al, *Machine Learning with Personal Data: Is Data Protection Law Smart Enough to Meet the Challenge?* International Data Privacy Law, Vol.7:1, p.1–2(2017).

㉗ *See* John Faber, *How to Future-Proof Your Search Ranking*, at https://www.chapterthree.com/blog/how–to–future–proof–your–search–ranking (Last visited on June 20, 2021).

㉘ *See* Danny Sullivan, *Google Uses Rank Brain for Every Search, Impacts Rankings of "Lots" of Them*, at https://searchengineland.com/google–loves–rankbrain–uses–for–every–search–252526 (Last visited on June 20, 2021).

（三）算法解释权的时间要求

就算法解释的时间要求而言，事前解释的意义有限。事前解释只能是一般解释。个体往往对于告知不太在意或难以理解，而机器学习等算法又可能时时发生变化，在这样的背景下，对算法进行事前解释，无助于个体理解算法和作出理性决策。对于个体而言，事前解释最多等同于告知个体“存在自动化算法决策”。这种解释既无法向个体告知演进后的算法规则，也无法针对具体个体的情况，向个体解释针对其个人的具体算法决策如何作出。[29]

同时，事后解释也面临难题。一方面，事后解释是在算法决策已经作出之后进行。自动化决策者无需提前告知个体其运行机制，那么个人对自动化决策与个人信息处理的选择权将形同虚设。缺乏知情与了解，个人的选择权与决定权可能成为一种任意行使的情绪性主张，缺乏理性基础。此外，在一些具有风险性的算法决策中，事后解释可能面临更大的问题。因为此类算法决策可能给个体带来重大风险，需要个体采取措施进行积极预防，事后解释显然无法帮助个体理解和防范此类风险。[30]

另一方面，事后解释应当以体系性解释作为要求，还是应当以个案性解释作为要求，也是一个两难问题。在机器学习等算法中，算法可能在经过大数据训练后发生变化，因此，算法的事后解释既可以要求自动化决策者对算法进行一般解释，也可以针对决策时的算法系统进行个案解释。如果将事后解释的要求界定为体系性解释，那么企业等算法主体需要对算法的演化进行追踪了解，将决策时的算法告知用户即可。其好处是企业等算法主体的负担相对合理，但问题在于，个体对算法决策的困惑可能很难消除。而如果将事后解释的要求界定为个案解释，则企业不但需要追踪算法的演化，还需要针对成千上万用户进行个性化的解释。此类解释虽然有利于用户消除疑虑，但也会给企业等算法主体施加不现实的负担。

（四）算法解释权的方式要求

在解释方式要求上，机器解释与人工解释也难以作统一要求。如果将人工解释作为强制性与统一性要求，并不现实。这意味着在个人信息的收集、处理、申诉等各个阶段，使用算法的企业都需要单独通知个人，并在个人行使访问权、更正权、删除权、拒绝自动化决策权等权利时都设置人工服务。在用户量有限的医疗算法、信贷算法等场景下，医院或大型银行或许可以满足这一要求，因为此类场景更类似于一对一或一对多服务。但在其他场景下，人工解释会给某些算法主体带来难以承担的压力。有的互联网与科技企业的用户超过数亿，每天存在无数的自动化处理。如果人工解释或服务是强制性要求，那么只要有千分之一的用户提起算法解释请求，企业就将不堪重负。

事实上，即使对算法解释权最为推崇的欧盟，也将很多自动化决策视为符合法定要

---

[29] *See* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, U.C. Davis Law Review, Vol.51:653, p.658–659(2017).

[30] *See* Jessica M. Eaglin, *Constructing Recidivism Risk*, Emory Law Journal, Vol.67:59, p.67–68(2017).

求的个人信息处理，毋需人工介入或人工解释。如，欧盟第 29 条数据保护工作组发布的《个人自动化决策和用户画像指南》认为，如果企业海量的自动化处理是企业所必须，那么此类处理可以被视为《一般数据保护条例》第 6(1) (b) 所规定的“为履行合同而必需”这一处理的合法性基础。对于《一般数据保护条例》规定的知情权、访问权、更正权以及拒绝自动化决策权，此类处理将不再适用。欧盟第 29 条数据保护工作组以招聘中的自动化算法为例，指出如果某企业工作很受欢迎，收到了数以万计的申请，那么企业可以合法使用纯粹的自动化决策筛除不相关的申请，以确定合适的候选人。[31]

但是如果完全不作要求，允许企业以纯机器解释的方法来进行算法解释，那么算法解释制度设计的初衷就可能大打折扣。在一定程度上，算法解释制度就是为了引入人类的解释说明机制，消除个人对算法的疑虑。如果允许以机器解释机器决策，则算法解释机制毋宁说是另一种自动化决策，未必能够消除个人疑虑。特别是当算法对个人的“权益造成重大影响”，而个人又对机器的算法解释感到困惑与质疑，希望有人工解释介入时，此时企业如果仍然未能提供人工解释的选项，则所谓的算法解释制度将形同虚设。

四、算法解释权的原理分析

解决算法解释权面临的法律解释争议与法律适用困境，需要回到算法解释权的基本原理上来。从基本原理看，现行算法解释权建立在算法个体控制论的基础之上，但这一理论面临种种困境。应当超越算法个体控制论，以沟通信任的基本原理重构算法解释权。从性质来看，这也意味着算法解释权应当被视为一种程序性权利，而非一种刚性的实体性权利。

（一）算法个体控制论的困境

在原理层面，算法解释权的提出与制定，与个人信息控制权理论密切相关。20 世纪 60 年代，美国学者阿兰·威斯丁（Alan Westin）首先提出了个人对于自身信息控制的权利，以回应传统隐私权在信息技术时代保护个人权益的不足。[32] 此后，这一个人信息控制权或个人信息自决权广泛传播，并在制度层面形成了以“公平信息实践”（fair information practice）为基础的个人信息保护制度。各国的个人信息保护制度虽然各有不同，但大多赋予个人以信息访问权、纠正权、删除权等控制性权利，同时要求信息处理者承担一系列责任。算法解释权作为一种新型权利，正是访问权、纠正权、删除权等传统个人信息控制权的进一步拓展。算法解释权期望通过个体对算法的知情、拒绝与控制，最大限度地保护公民在大数据与自动化决策中的权益。[33]

但将算法解释权视为绝对化的控制性权利，存在如下几方面的问题。首先，个人信

---

[31] *See* Article 29 Working Party, *supra* note 17.

[32] *See* Alan Westin, *Privacy and Freedom*, New York: Atheneum, 1967, p.7.

[33] *See* Margot E. Kaminski, *The Right to Explanation, Explained*, Berkeley Technology Law Journal, Vol.34:189, p.190(2019).

息控制权本身存在可行性困境。虽然法律在个人信息收集、储存、流转的各个环节都赋予了个体以知情选择权，但个人很难通过“告知—同意”的方式来作出理性判断，以至于这种权利在很大程度上仍然是停留在“纸面上的权利”。[34] 同时，解决这种困境，也无法简单通过强化个人选择权来完成。为了应对“告知—同意”的失效，有的方案要求信息处理者简化隐私政策，要求信息控制者制定简短易懂的隐私政策[35]，有的方案要求信息控制者都以“选择加入（opt-in）”[36]等方式获取消费者同意。但这些改良版的方案却不仅无法解决个人控制的问题，反而可能带来更多问题。

在算法问题上，个人控制权面临更多问题。算法自动化决策比个人信息的收集、储存与流转更加复杂，更加具有黑箱属性与不可预测性。[37] 以近年来兴起的可解释人工智能（XAI）项目为例，这一项目的最初目标被定义为实现人工智能的可解释性（explainability）。但面对算法的黑箱属性与机器的自我学习特征，很多学者认为，算法全面解释不可行或难度太大，应以可理解性（interpretability）作为目标，即“有能力用可理解的术语向人类表达”[38]，而非对算法进行全盘解释。理解算法对于专家已是非常困难，对于普通个人来说就更是如此。[39]

其次，个人控制权存在正当性困境。在很多自动化决策的场景中，自动化决策者与个人之间的关系具有互惠合作性，而不只具有对抗性，这就使得个人的算法解释权与控制权难以具备绝对化权利的正当性基础。一般而言，当法律关系中的两方是对抗防范关系，同时一方具有强势地位或行使公权力，此时作为弱势防范的一方，有权要求对方对相关行为进行说明，并在没有法律授权的情形下获得个人同意。但在算法自动化决策中，很多企业的算法决策是为了向个体提供更有效的信息、更优惠的价格和更好的服务。对于企业而言，只有为消费者提供性价比更高的产品，才更有利于企业长期发展。如果法律关系中的双方形成的是互惠关系，那么赋予一方以更多的解释权，不仅没有必要，而且会给双方的合作关系带来不必要的烦扰。

当然，自动化决策者与个人之间也经常具有对抗性。[40] 如，在网约车与电商商品推荐的场景中，企业可能利用算法进行“大数据杀熟”；在搜索引擎的场景中，企业可能利

---

[34] *See* Kenneth A. Bamberger & Deirdre K. Mulligan, *Privacy on the Books and on the Ground*, Stanford Law Review, Vol.63:247, p.247(2011).

[35] 如，《一般数据保护条例》第 12 条第 1 款规定，当控制者试图获取用户的同意时，“控制者应当以一种简洁、透明、易懂和容易获取的形式，以清晰和平白的语言来提供”相应信息。

[36] 对于“选择加入”与“选择退出”的分析，可参见 Hans Degryse & Jan Bouckaert, *Opt in Versus Opt Out: A Free-Entry Analysis of Privacy Policies*, at https://ssrn.com/abstract=939511(Last visited on June 20, 2021)。

[37] 参见林洹民：《自动决策算法的法律规制：以数据活动顾问为核心的二元监管路径》，载《法律科学》2019 年第 3 期，第 48 页。

[38] Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, Nature Machine Intelligence, Vol.1:206, p.206(2018).

[39] 参见沈伟伟：《算法透明原则的迷思——算法规制理论的批判》，载《环球法律评论》2019 年第 6 期，第 25-26 页。

[40] *See* Gerhard Wagner & Horst Eidenmueller, *Down by Algorithms? Siphoning Rents, Exploiting Biases, and Shaping Preferences: Regulating the Dark Side of Personalized Transactions*, University of Chicago Law Review, Vol.86:581, p.609(2019).

用搜索算法而诱使用户进行某些不必要的消费。现实中已经出现了大量的企业支配消费者和榨取消费者剩余的情形。[41] 在此类不对等的对抗防范关系中，赋予个体以拒绝算法自动化处理权与解释权，的确具有正当性与必要性。但这种反对权与解释权的行使应以企业存在不合理支配为前提，以消除企业对个人的欺诈与不公平对待为目的。[42] 对于不具有对抗性或具有互惠合作性的关系，赋予刚性的权利则并无必要，也不利于双方合作关系的形成。[43]

最后，市场本身也具有调整能力，在企业算法决策中，并非所有不合理的算法问题都需要国家力量的介入。当某一领域的竞争较为充分，特别是已经形成市场良性竞争的前提下，当某企业利用算法对个人设置过高的价格或提供低水平服务，这家企业就很可能会被消费者抛弃，被市场淘汰。因此，算法解释权应当更多针对那些对消费者产生重大影响的自动化决策，或者针对具有市场支配地位的企业。前者可以使政府力量集中，避免规制中经常出现的规制失灵、选择性规制、规制俘获等问题。[44] 后者则可以避免某些垄断企业利用算法和消费者的个人信息而滥用市场支配地位，支配消费者。[45]

（二）抛弃算法解释权？

算法解释权既然面临种种困境，那是否应当抛弃算法解释权的进路，以算法规制的进路作为替代？如，政府可以对算法进行直接监管，要求企业在某些敏感领域对算法进行审查与风险评估；[46] 政府也可以通过消费者保护法等框架来监管算法，要求企业的算法决策不得欺诈消费者；政府还可以用企业自我监管的方式对待企业算法。这些模式与算法解释权的区别在于，这些模式主要从外部对算法进行监管，都没有赋予个体以主张算法解释的权利。

算法外部监管具有很多优势，例如监管机构的专业能力较强，对于算法的理解能力高于普通个人；监管机构的人力物力远非个人能比，因此其调查与执法能力强于个人；监管机构通过对算法的监管与风险评估，可以保护所有使用该算法的个人，而不仅仅是某个个人。但彻底放弃算法解释权的进路，代之以算法的外部监管，并不是合理的法律与制度选择。

首先，算法的外部监管面临多重挑战。监管机构虽然专业能力、执法能力较强，但监管机构的数量和人员都有限度，发现问题能力较弱。尤其是在算法这类科技专业性较强

---

[41] 参见张凌寒：《算法权力的兴起、异化及法律规制》，载《法商研究》2019 年第 4 期，第 65 页；周辉：《算法权力及其规制》，载《法制与社会发展》2019 年第 6 期，第 113–114 页。

[42] 参见李成：《人工智能歧视的法律治理》，载《中国法学》2021 年第 2 期，第 128 页。

[43] 参见丁晓东：《社会法概念反思：社会法的实用主义界定与核心命题》，载《环球法律评论》2021 年第 3 期，第 94 页。

[44] *See* Michael Levine & Jennifer L. Forrence, *Regulatory Capture, Public Interest, and the Public Agenda: Toward a Synthesis*, Journal of Law Economics & Organization, Vol.6:167, p.167–168(1990).

[45] *See* James C. Cooper, *Privacy and Antitrust: Underpants Gnomes, the First Amendment, and Subjectivity*, George Mason Law Review, Vol.20:1129, p.1129(2013).

[46] *See* Bryan Casey, Ashkon Farhangi & Roland Vogl, *supra* note 8,144.

的领域，一般监管机构的通用专业能力往往难以应付。[47] 面对快速变化的科技问题，监管机构往往难以进行知识更新，赶上科技的发展。[48] 此外，监管往往依赖全有全无的规则或命令，对于场景多元、权益多变的算法自动化决策来说，这类规则或命令也未必完全适合监管算法。[49] 可见，政府对算法的外部监管具有监管能力和监管手段方面的局限性，完全依赖政府外部监管效果不佳。正如有的学者所言，“如果指望国家单独解决算法问责问题，那就找错地方了”[50]。

其次，在算法外部监管面临困境的方面，算法解释权的合理行使恰巧可以弥补其不足。面对算法决策，个体虽然认知能力较弱，也一般倾向于息事宁人，不愿意介入法律诉讼，但用户是算法决策的最直接影响对象，算法的变化或更改，用户的感受往往最为直接深刻。即使其中有很小比例的用户发现问题，提起申诉或诉讼，也可以较快发现问题。因此，个人对于算法的控制权虽然难以行使，但赋予个人对于算法的有限解释权，通过个人发现算法存在的问题，有利于对算法进行监督和合作治理[51]。

最后，算法解释权不仅可以扮演一种工具性权利，也可以成为一种价值本身。在日常交往与法律活动中，受到影响特别是受到伤害的一方寻求解释，这对于维护个体自治与尊严价值具有重要意义。法律心理学的成果表明，法律程序的一个重要价值就在于使相关主体得到倾听，相关疑虑得到解释。通过获得解释与参与相关沟通，个人可以在法律程序中获得一种被尊重的感受。[52] 在算法自动化决策中，个体主张某种程度的算法解释权，也可以被视为一种直觉性的个体尊严需求。[53] 算法解释权的合理行使，有利于个人自主性的实现。[54]

（三）迈向沟通信任的算法解释权

为了避免以个人控制权为基础的算法解释权带来的困境，同时弥补算法外部监管的不足，算法解释权的基本原则应当进行重构。既然自动化决策者与个人既存在合作互惠关系，又存在对抗防范关系，那么算法解释权就应当以促进双方的合作互惠为目标，以消除个人的疑虑和防范自动化决策者的不合理支配为目标。[55] 也就是说，算法解释权应当

---

[47] *See* Emily S. Bremer, *Private Complements to Public Governance*, Missouri Law Review, Vol.81:1115, p.1116(2016).

[48] *See* Meg Leta Jones, *Does Technology Drive Law? The Dilemma of Technological Exceptionalism in Cyberlaw*, Journal of Law, Technology & Policy, Vol.2018:249, p.277(2018).

[49] *See* Dennis D. Hirsch, *The Law and Policy of Online Privacy: Regulation Self-Regulation, or Co-Regulation?,* Seattle University Law Review, Vol.34:439, p.458–459(2011).

[50] Sonya K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, UCLA Law Review, Vol.66:54, p.61(2019).

[51] *See* Margot E. Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, Southern Califonia Law Review, Vol.92:1529, p.1529(2019).

[52] *See* Dan M. Kahan, *The Economics Micro Behavioral, and Political of "Subsequent Remedial Measures" Evidence*, at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=15618439(Last visited on Sept. 11,2021).

[53] *See* Andrew D. Selbst & Solon Barocas, *supra* note 8,1085.

[54] 参见陈姿含：《人工智能算法中的法律主体性危机》，载《法律科学》2019 年第 4 期，第 40 页。

[55] 参见郑智航：《人工智能算法的伦理危机与法律规制》，载《法律科学》2021 年第 1 期，第 14–15 页。

建立在沟通信任的原则之上。[56]

在个人信息保护的基础理论研究中，以信任原则建构信息隐私保护，已经涌现了一大批研究成果。这些理论的共识是，以个人控制论为基础的个人信息保护已经难以承担网络时代的信息隐私保护，信息处理者应当超越一次性的告知同意，构建可信赖的信息处理机制，赢得信息主体的信任。如，阿里·瓦尔德曼（Ari Waldman）认为，信息隐私的本质应当是信任，而非关于个人控制，以信任关系界定信息隐私，更有利于在互联网与大数据时代保护个人权益。[57]

此外，杰克·巴尔金（Jack Balkin）等学者提出的信息信义义务理论，也与信任理论密切相连。针对互联网时代个人与信息处理者之间的信息能力不平等，个人对于信息处理者的依赖，平台的专业信息处理能力，以及二者之间可能形成的信任关系，[58]巴尔金认为应当对平台施加信义义务，要求平台承担对个人的保密义务（duty of confidentiality）、谨慎义务（duty of care）和忠诚义务（duty of loyalty），[59]以建构一种基于值得信任的平台个人信息保护。虽然有学者对巴尔金的理论提出质疑，认为这种信义义务理论不足以保护个人的信息隐私，[60]但作为一种对个人控制模式的批判与替代，信息信义义务仍然具有强大的生命力，也得到了很多权威学者的支持。[61]

在算法自动化决策问题上，以信任原则重构算法解释权，比个人信息保护中的其他议题更为必要。在算法自动化决策中，个人所面临专业化壁垒更高、信息更不对称、不确定性更高。对于这种关系，法律更难通过告知同意的方式进行规制。如果说个人信息保护中有的议题更类似于消费者法保护，存在一定的能力与信息不对称；那么算法自动化决策则类似于医患关系，双方的能力与信息不对称程度更高。在这样一种关系中，法律经常要求医生对病患承担信义义务，建构一种基于信任而非简单同意机制的医患关系。[62]对于算法自动化决策，也应当更多以信任原则和信义法的框架重构算法解释权。

（四）算法解释权的程序性特征

从沟通信任的原则出发，可以对算法解释权的性质进行重新界定。既不应将算法解释权视为一种透明性要求，也不应将其视为一种静态、孤立、绝对化的实体性权利。相反，

---

[56] 算法规制中的信任问题，参见苏宇：《算法规制的谱系》，载《中国法学》2020年第3期，第165页。

[57] *See* Ari Ezra Waldman, *Privacy as Trust: Sharing Personal Information in the Twenty-First Century*, University of Miami Law Review, Vol.69:559, p.559(2015).

[58] *See* Jack M. Balkin, *Information Fiduciaries and the First Amendment*, U.C. Davis Law Review, Vol.49:1183, p.1222(2016).

[59] *Id.*, 1206–1208.

[60] *See* Lina M. Khan & David E. Pozen, *A Skeptical View of Information Fiduciaries*, Harvard Law Review, Vol.133:497, p.497 (2019).

[61] *See* Jonathan Zittrain, *Engineering an Election: Digital Gerrymandering Poses a Threat to Democracy*, at https://harvardlawreview.org/2014/06/engineering-an-election/(Last visited on Sept.11,2021); Neil Richards, *Intellectual Privacy: Rethinking Civil Liberties in the Digital Age*, Oxford University Press, 2015, p.168; Jack M. Balkin, *The Fiduciary Model of Privacy*, Harvard Law Review, Vol.134:11, p.33(2020).

[62] *See* Onora O'Neill, *Some Limits of Informed Consent*, Journal of Medical Ethics, Vol.29:4, p.6–7(2003).

算法解释权应当被视为一种动态、沟通、相对性的程序性权利，[63] 因为信任本身就是持续性、关系性、程度性的。

首先，信任本身是一个过程，无法通过一次性授权而永久性建构，[64] 这使得算法解释权更类似一种动态性权利。正如有的学者指出，信任本身就是一种贝叶斯态度，一方对另一方的信任常常取决于先前事件的变化而变化，而非取决于最初双方的合意。[65] 现代合同法研究也表明，所有合同都只能是不完备合同（incomplete contract），即使在最经典的一般合同关系中，双方也不可能对所有事项进行约定，而需要双方信任关系的持续性构建。[66] 在算法信任关系的建构中，持续性问题更为突出，更需要消除双方或一方的疑虑。如此一来，算法解释权就不能够限定在某一个时间点对于个人的一次性解释。这种解释即使在当时赢得了个体的信任，也可能因为时间的演化而受到个体怀疑，需要后续的更多解释。

其次，信任需要在关系中进行建构，而不能通过孤立的权利行使来获取，这使得算法解释权更类似沟通性权利。目前，我国和欧盟的法律赋予了个体在算法自动化决策中的反对权与申诉权，如果行使恰当，这些权利可以充当沟通与信任的桥梁，既消除个体的疑虑，同时也可以成为企业等信息主体倾听消费者、改进服务的渠道。但如果行使不当，此类权利也可能成为进一步破坏信任的壁垒。个人可能仅仅得到一种机械性的回应，不仅没有消除疑虑，反而产生更多的怀疑；而企业则可能将回应机制视为一种负担和应付，对算法解释消极应对。因此，孤立和形式主义的权利行使无助于信任的建构，应以关系沟通的立场看待算法解释权。

最后，信任是程度性而非绝对性的，这使得算法解释权更接近于一种相对性权利。就信任而言，即使是最亲密的关系，也不可能实现百分之一百的信任。在商业算法决策等场景下，信任关系更不可能达到私人之间亲密关系中的信任程度，也毋需达到这种程度。因此，算法解释权的行使不应追求绝对信任。目前，我国和域外国家的法律已经作出了某些规定，例如将对个体权益造成重大影响作为限定条件，此类规定有利于以信任原则重构算法解释权。此外，算法解释权中的信任问题还应考虑市场竞争、算法决策的应用领域。例如对于市场竞争较为激烈，用户选择权较为充分的领域，就不应对双方信任程度作过高要求。[67] 相反，对于公共事业领域的算法，或者那些具有一定市场垄断、具有公共基础设施性质的企业，则应要求或引导企业获取个体更高程度的信任。在算法解

---

[63] *See* Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, Washington Law Review, Vol.89:1, p.16–18(2014)；陈景辉：《算法的法律性质：言论、商业秘密还是正当程序？》，载《比较法研究》2020 年第 2 期，第 131–132 页；李晓辉：《算法商业秘密与算法正义》，载《比较法研究》2021 年第 3 期，第 121 页。

[64] *See* Onora O'Neill, *Autonomy and Trust in Bioethics*, Cambridge University Press, 2002, p.44.

[65] *See* Russell Hardin, *Trust and Trustworthiness*, Russell Sage Foundation, 2002, p.113–114.

[66] *See* Ian R. Macneil, *Relational Contract Theory: Challenges and Queries*, Northwestern University Law Review, Vol.94:877, p.877(1999).

[67] 值得注意的是，消费者也可能利用算法进行选择。*See* Michal S. Gal & Niva Elkin-Koren, *Algorithmic Consumers*, at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2876201(Last visited on Sept. 11,2021).

释权的规则设定上，国家“有必要更多地采取软法方式，而不是简单地提高硬法的惩戒力度”。[68]

总之，算法解释权的权利主张不应以绝对透明为目标。绝对透明的要求可以永无止境，信息的无限度披露也只会引起信息过载与选择疲劳。算法解释权也不应是绝对性的。个人有向自动化决策者提出算法解释或者反对算法决策的正当程序权利，自动化决策者也应建立回应这类个人诉求的机制。但此类权利主张应更类似于一种程序性权利，而自动化决策者的回应机制也更应被视为一种沟通机制或客服机制。通过正当程序权利的行使与回应，算法解释权可以成为建立算法信任与科技向善的桥梁。

五、算法解释权的制度重构

从算法解释权的原理出发，可以重新建构算法解释权制度。算法解释权可以根据其权利内容、解释程度、解释时间、解释方式不同而分为一般性主张与限定性主张、系统性解释与个案性解释、事前解释与事后解释、机器解释与人工解释，而且每种分类均存在法律解释层面与法律有效性层面的难题。要回应这些难题，可以结合算法解释权的沟通信任原则与程序性性质，对算法解释权的内容、程度、时间与方式作不同类别的要求。

首先，就权利内容而言，算法解释权不应被视为一种内容边界完全清晰的权利。无论是一般性主张还是限定性主张，其权利主张都是程序性的，其实体性边界应该根据具体场景下个人与算法决策者之间的沟通信任关系而确定。总体而言，当算法决策所处的场景具有严重的信任危机，且难以通过市场竞争改善算法时，应当赋予个体一般性的算法解释权主张，而不仅仅是拒绝算法自动化决策的权利。同时，个体主张算法说明解释的权利也不应设置前提。例如，当人们对外卖算法系统存在普遍不信任时，或者对网约车的算法派单感到困惑时，应赋予个体以算法解释权，这类一般性主张将有利于督促自动化决策者对算法进行改善，重新赢取社会与用户的信任。相反，当算法决策所处的场景本身具有相对较高的信任度，或者该行业具有良性互动的竞争机制，则此时应将算法解释权限定为某些特定的权利主张。一部分群体可能对自动化决策本身存在怀疑，当自动化决策对他们产生重大影响，赋予他们以拒绝自动化处理的权利或算法说明解释权，有利于进一步促进二者之间的信任。

其次，就解释程度而言，一方面应要求企业充分了解和掌握算法的系统性解释。现代互联网企业和银行的算法往往“由众多工程师设计组成的复杂公式决定”，企业或规制机构掌握其算法的整体运行机制，这是算法解释权有效实施的前提。[69] 为了实现这一

---

[68] 参见季卫东：《人工智能开发的理念、法律以及政策》，载《东方法学》2019 年第 5 期，第 4 页；衣俊霖：《数字孪生时代的法律与问责——通过技术标准透视算法黑箱》，载《东方法学》2021 年第 4 期，第 77 页。

[69] *See* Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, 2016, p.33.

目标，企业或运用算法决策的规制机构就应打通其内部壁垒，在企业内部树立算法伦理与合规实践，[70] 真正将企业向消费者告知的隐私政策和对用户的回应落实到企业的每一个环节。换句话说，企业需要不断进行内部的自我监管，既需要在事前解释环节就对算法设计进行内部沟通，也需要在事后解释中倒查企业内部的算法运作机制，保持对算法的内部动态监管。[71] 另一方面，可以根据算法所处的领域特征、影响性不同而要求不同程度和类别的算法个案解释。对于那些用户具有较多选择权、个案影响一般的算法决策，应允许自动化决策者自身设定个案解释的规则。对于企业应用算法进行不涉及意识形态与公共伦理的娱乐资讯的分发，应允许企业在个案解释中进行自我规制，避免国家过多的强制性监管。但对于具有公共性影响的算法以及对个体造成重大影响特别是实际伤害的算法决策，国家则应强制自动化决策者进行个案解释，以保护公共利益与个人的核心权益。如，在医疗场景下利用机器人进行手术，此时应当充分保障患者对于算法的个案知情权，患者不仅有权知晓手术机器人算法决策的一般规则，而且应当有权了解该机器人对其个人进行手术的决策机制、过程与效果。[72] 又如，在涉及外卖骑手人身安全、卡车司机罚款或者自动驾驶汽车安全的算法决策中，个人应当有权在这些情形中提出算法解释的请求。

再次，就解释时间要求而言，在告知环节可以要求或倡导企业在事前进行模糊性解释。[73] 由于算法的自我演化与不确定性，算法向公众的告知可以描述算法的整体运行，让用户或消费者感知算法的具体运行情况，帮助个体作出更佳决策，[74] 而非对算法的所有参数和所有细节进行描述。例如，网约车的算法可以向用户告知，其算法是否会将性别、收入、年龄、高峰期的车辆运行等情况纳入算法规则的解释说明中，但没有必要对某一个参数如何影响算法进行具体告知。[75] 此类描述不仅可以在一定程度上克服算法的不确定性问题，而且有利于克服系统性解释的针对性不强、侵犯商业秘密、被人利用等难题。因为此类解释可以在一定程度上消除用户较为普遍的疑虑，同时，由于此类描述较为原则，此类解释也不会对企业的商业秘密造成严重侵害，一些钻空子的个人与企业也无法“算计”算法。

最后，就解释方式而言，自动化决策者可以根据自身能力、行业特征、用户数量与算

---

[70] *See* Neil M. Richards & Jonathan H. King, *Big Data Ethics*, Wake Forest Law Review, Vol.49:393, p.408–409(2014).

[71] *See* Michael Guihot et al., *Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence*, Vanderbilt Journal of Entertainment and Technology, Vol.20:385, p.385–427(2017); 金梦：《立法伦理与算法正义——算法主体行为的法律规制》，载《政法论坛》2021 年第 1 期，第 29 页。

[72] *See* W. Nicholson Price II, *Regulating Black-Box Medicine*, Michigan Law Review, Vol.116:421, p.465–471(2017).

[73] 企业隐私政策的告知功能也有利于企业的自我规制。*See* Peter P. Swire, *The Surprising Virtues of the New Financial Privacy Law*, Minnesota Law Review, Vol.86:1263, p.1316(2002).

[74] *See* Lilian Edwards & Michael Veale, *Enslaving the Algorithm: From a “Right to an Explanation” to a “Right to Better Decisions”?*, IEEE Security & Privacy, Vol.16:46, p.46–54(2018).

[75] 这一进路类似于有的学者所提出的“反事实解释”。*See* Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*, Harvard Journal of Law & Technology, Vol.31:841, p.841(2018).

法影响程度而选择人工解释或机器解释。在理念层面，人工解释具有一定的优先性，专业高效的人工解释更有利于信任的建立。同时，对于一些对个人或社会造成重大影响的算法，人工解释应成为兜底性要求，以便消费者或监管机构对企业或算法运用的主体进行算法监督。[76] 但对于其他并未产生重大影响、用户数量庞大的算法，要求企业或算法主体对所有个体都采取人工解释，并不现实。如果强行施加此类要求，其结果可能反而是企业关闭所有的解释与沟通渠道。因此，应当允许企业在一般情况下运用机器解释或自动化客服。机器解释如果运用得当，例如在事先充分设想场景，为用户提供各类“反事实解释”或针对性解释，[77] 则机器解释也能有效起到提供“有效性信息”、沟通解惑的功能。此外，还可以鼓励企业或算法主体建立机器与人工的混合解释机制。机器解释与人工解释的混合使用，将有助于减小自动化决策者的现实压力，更有效推动算法解释权的落地。

六、结　语

算法解释权是个人信息被保护权中的一种权利。对于是否应当将个人信息受保护权上升为法定权利，学界本身就已经产生很多争论。[78] 支持者认为，以个人信息受保护权制约信息处理者，有利于维护公民权益。[79] 反对者则认为，对于个人与信息处理者之间的关系，更应以私法自治的框架加以调整，个人信息受保护权只是个人合同性权利或者消费者权利中的一种。[80] 随着我国《个人信息保护法》的制定与生效，这一权利已经被我国实证法认可。但就法律解释与学术研究而言，这一权利仍有待进一步讨论。

作为个人信息被保护权的一种，算法解释权不仅具有个人信息保护中的一般争议性问题，而且更具有特殊性。算法具有黑箱性、不确定性、复杂性，因此相比个人信息的访问权、更正权、删除权等传统权利，算法解释权更具有争议。算法解释权是否可以成为一种一般性权利，还是仅仅指拒绝算法决策的特定性权利？算法解释权可以主张对算法进行系统解释还是个案解释，事前解释还是事后解释？机器解释还是人工解释？此类问题已经引起法律解释争议与法律适用困境。

本文对算法解释权进行原理反思，指出应当以信任关系重构算法解释权的基本原理。算法解释权的目的不是为了实现“没有意义的透明”[81]，或者建立个体对于算法的绝

---

[76] *See* David C. Vladeck, *Machines without Principals: Liability Rules and Artificial Intelligence*, Washington Law Review, Vol.89:117, p.150(2014).

[77] *See* Sandra Wachter, Brent Mittelstadt & Chris Russell, *supra* note 76, 841.

[78] 参见周汉华：《个人信息保护的法律定位》，载《法商研究》2020 年第 3 期，第 44 页；王锡锌：《个人信息国家保护义务及展开》，载《中国法学》2021 年第 1 期，第 145 页。

[79] *See* Paul M. Schwartz, *Global Data Privacy: The E.U. Way*, New York University Law Review, Vol.94:771, p.771(2019).

[80] *See* Omri Ben-Shahar, *Contracting Over Privacy: Introduction*, Journal of Legal Studies, Vol.45:1, p.11(2016).

[81] Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why A “Right to an Explanation” is Probably not the Remedy You Are Looking for*, Duke Law & Technology Review, Vol.16:18, p.23(2017).

对性控制。相反，算法解释权是为了在信息处理关系中搭建桥梁，在个人与自动化决策者之间构建信任、消除误解、减少疑虑。从信任沟通原则出发，本文主张将算法解释权视为一种程序性权利，而非实体性权利。[82] 在制度建构层面，应当根据算法自动化决策所位于的行业领域、市场化程度、个案影响、企业能力而对算法解释权的内容、程度、时间和方式作不同要求。总之，算法解释权应当成为个体与自动化决策者之间的信任沟通机制，促进“负责任算法”[83] 的实现。

Abstract: Algorithmic automatic decision-making based on personal information is becoming more and more common in contemporary society. The *Personal Information Protection Law of China* and relevant domestic and foreign laws all introduce the right to algorithm interpretation to deal with it. However, due to the black box nature, uncertainty and complexity of the algorithm, there are interpretation disputes and application dilemmas as to whether the right to algorithm interpretation should be understood as general right or limited right, systematic interpretation or case interpretation, ex ante interpretation or ex post interpretation, and manual interpretation or machine interpretation. At the principle level, this controversy and dilemma stems from individual algorithm control. The right to algorithm interpretation should be reconstructed based on the principle of communication and trust, and the nature of the right to algorithm interpretation should be regarded as a procedural right rather than a substantive right. The institution of the algorithm interpretation right should also be reconstructed. The content, degree, time and method of algorithm explanation right should be regulated according to the industry field, marketization degree, case influence and enterprise ability.

（责任编辑：强梅梅）

[82] 信任沟通原则也开始逐渐为我国监管机构采用。例如由国家互联网信息办公室等九部委制定的《关于加强互联网信息服务算法综合治理的指导意见》第 13 条规定：“做好算法结果解释，畅通投诉通道，消除社会疑虑，推动算法健康发展。”

[83] Joshua A. Kroll et al., *Accountable Algorithms*, University Pennsylvania Law Review, Vol.165:633, p.633(2017).