算法伤害和解释权

邵国松 黄琪

摘要

算法对个人和社会的影响日益明显，同时引起了一系列围绕算法决策不公、歧视与不透明的隐忧。旨在使算法透明化的解释权被认为是应对算法伤害的有效措施。然而，解释权在现有的法律资源中难以证成，在实践中也无法有效解决算法对个体造成的损害。本文认为，无需另创解释权，GDPR的数据擦除权、数据携带权、数据保护影响评估等权利机制与范式要求便可以使算法更合理和更负责。

关键词

算法决策、算法伤害、解释权、数据保护通例（GDPR）

作者简介

邵国松，上海交通大学媒体与传播学院副院长、教授、博导，电子邮箱：gshao@sjtu.edu.cn。

黄琪，安徽省明光市人民法院法官助理，电子邮箱：huangqi1412@126.com。

本文系国家社会科学基金一般项目“大数据时代个人数据法律保护研究”（项目编号：15BXW030）的研究性成果。

Algorithmic Harms and the Right to Explanation

SHAO Guosong HUANG Qi

Abstract

The growing influence of algorithm agents on individuals and the society is becoming obvious; meanwhile, they have caused a series of public concerns such as the unfairness, discrimination, and opaqueness of algorithmic decisions. The right to explanation purporting to make the algorithm decision-making transparent is thus considered as an effective measure to deal with algorithmic harms. However, the right to explanation hardly can be proved in the existing legal framework. Furthermore, it cannot effectively solve the algorithmic harms to individuals in reality. We argue that it is not necessary to create a new right of explanation since such GDPR rules as the right to erasure, right to data portability, and data protection impact assessment, can make algorithm decision-makings more reasonable and responsible.

Keywords

Algorithmic decision-making, Algorithmic harm, Right to explanation, General Data Protection Regulation (GDPR)

Authors

Shao Guosong is Associate Dean and Professor of School of Media and Communication at Shanghai Jiao Tong University. Email: gshao@sjtu.edu.cn.

Huang Qi is a judge assistant of Mingguang people's Court of Anhui Province. Email: huangqi1412@126.com.

This paper is the research achievement of the general project "Research on the legal protection of personal data in the era of big data" (Project No.: 15bxw030) of the National Social Science Foundation of China.

自人类进入互联网时代尤其是时下的大数据时代后，算法日益规制我们的生活。无论在商业还是公共领域，商业或政府机构日益依赖算法进行决策，或为其决策提供算法支持。在日常生活领域，人们高度依赖的GPS导航、搜索引擎，社交平台之信息分享、购物网站之商品推荐等均为算法决策的结果。算法已深度介入了人类的生活，我们所享受到的各种网络红利基本上是算法运行的结果；但如果缺乏合理规制，算法也有可能构成对我们权益的损害。

与人工决策相比，算法决策具有相对客观、高效、便利等特点。在公众看来，算法具有技术中立特性，仅依据预先设计的技术逻辑以及数据之间的关联性进行判断、决策。然而，算法并非我们想象的这般简单，它蕴含着设计者乃至使用者的伦理判断或特定的价值立场，许多算法在设计或使用之时就嵌入了不公与歧视性因素，有些算法更是被人为操控以实现特定目的（丁晓东，2017）。随着算法对社会的影响日增，人类开始担心其命运可能被一种难以理解、无法控制的系统所支配，这个系统无处不在但缺乏透明度，当它产生意想不到、破坏性、不公正或歧视性的结果时，人们并无明确的解决方法。

目前各国均在寻找算法给个人权利造成伤害的救济机制。一个核心问题是，如果算法做出不利于相对人（数据主体）的决策时，相对人是否有权了解其决策过程，又如何获得有效救济（张凌寒，2018a）。目前最常见的说法是使算法透明化，这通常意味着算法的可解释性。透明化在此指的是算法代码或数据的公开，可解释性指的是算法的概念或运行机制可向终端用户解释（New & Castro，2018）。二者概念不同，但在算法规制层面，算法透明化是为其可解释性服务的。

2018年5月，欧盟的《数据保护通例》（General Data Protection Regulation，以下简称GDPR）开始正式实施，该条例所包含的“解释权”（Right to an Explanation）被认为是一种打开黑箱、纠正算法、提升算法伤害可归责性的补救措施。然而，对于解释权是否存在、是否可以有效解决算法伤害等问题，学术界仍存诸多疑问，这也是本文要讨论的关键问题。

本文首先简要概括算法决策所带来的伤害，在此基础上详细讨论解释权的提出和疑义，之后寻求解释权之外的解决方案，最后对全文进行总结，并简要讨论解释权在我国的适用性问题。

一、算法伤害

（一）用户画像

用户画像指的是为了评估自然人的某些条件而对个人数据进行的自动化处理，特别是为了评估自然人的工作表现、经济状况、健康、个人偏好、兴趣、可靠性、行为方式、位置或行踪而进行的处理。算法并没有奇幻的“魔力”，没有数据的喂养，算法推送的信息是不可能主动精确匹配到个人的（例如，一个对股市毫无兴趣的人，资讯类APP是不可能经常给他推送股票信息的）。但用户在享受便利的同时，其信息隐私也在不知不觉中被侵犯。比如，要实现新闻的个性化推送，首先就必须广泛挖掘用户的浏览记录，对用户的喜好进行分析总结，绘制出用户画像，然后计算出每条新闻与用户画像的相似度，将相似度最高的新闻推荐给用户，从而达到个性化推送的效果（张潇潇，2017）。再比如，电商平台常常根据用户以往的购买记录，或相似用户的购买行为，或商品之间的关联，来精准推送用户所需的商品信息。一般说来，喜欢阅读何种新闻、购买何类商品、结交何种朋友等均属于用户的隐私范畴，而算法推荐以隐晦的方式告诉用户“你喜欢什么、你干了什么、你将要做什么，我全都知道”，不仅知晓用户已经存在的隐私，甚至也掌握了其未来的隐私。这样，无需见面沟通，无需入屋搜查，仅凭用户的“只言片语”和上网痕迹便可精准为其画像，在强大的算法技术面前，网络用户基本无隐私可言。

（二）算法黑箱

算法的运行有赖于复杂的机器学习能力，算法的决策过程具有天然的不透明性，因此算法经常被描述为“黑箱”。算法自动化决策不是根据人类易懂的规则做出的，不仅没有人工干预，而且常常缺少可解读的数学技术。更糟糕的是，算法所做出的决策可能是错误的、带偏见的甚至具有破坏性。例如，如果一个贷款申请人

因为其数据被错误输入，从而被信用评估算法拒绝贷款，他很难了解决策做出的机制而只能接受结果，更无法证明这个系统对他的非法歧视。相较于法律代码的可理解性，算法代码很不透明。我们对算法施加的权力没有清晰的认识，无法理解算法运行的原理。它们的决策原理隐藏在我们无法轻易理解的代码面纱中。算法根据数据模式不断迭代，这使得我们更难理解和预测（Perel & Elkin-Koren，2017）。此外，算法大多被追求利益最大化的商业组织所使用，而它们恰恰承受最小化的公开透明义务。因此，通过传统、被动的信息披露机制以便公开观察并生成适当的责任承担机制成为不可能。

（三）算法歧视

算法歧视常见诸于性别、种族、就业、贷款等方面。算法依据风险评估结果或相关分类，来决定是否提供一些机会，或将他人无需承担的成本强加于某客体。卡内基·梅隆大学的一项研究发现，谷歌公司开发的广告定位算法可能存在针对用户的性别歧视。具体来说，搜索年薪20万美元以上的行政职位，假冒男性用户组收到1852个广告，而假冒女性用户组仅收到318个广告（Tufekcl，2015）。还有研究发现，在美国劳动市场环境下，如果应聘者的名字听起来不像英美人常用的名字，则有可能成为算法歧视的对象（於兴中，2018）。与白人相比，美国黑人利用谷歌搜索时，搜索结果会出现“你是否有被捕记录”的询问（多梅尔，2016：103）。在进行图像分类时，由于用于机器学习的数据不足，没有足够适合标注的黑人图片，曾有算法系统将黑人夫妇识别为“猩猩”。在犯罪风险评估方面，美国部分法院所使用的风险评估算法COMPAS，被证明对黑人造成了系统性歧视（张凌寒，2018b）。在征信领域，未能充分积累个人声誉与信用评分的社会主体，更容易被排除在征信系统外（胡凌，2017）。

算法歧视之所以普遍存在，是因为算法的目的、设计理念、技术标准等无不渗透着设计者或使用者的主观选择与价值追求，设计者或使用者可能将自身的歧视或偏见嵌入算法决策系统。另外，数据质量本身也会影响算法决策的结果，因为算法是根据历史数据进行训练而生成的，存在复制不公平/歧视性历史记录的风险。数据的不正确、不完整可能造成算法决策所谓的“带病进出”现象。数据是社会现实的电子化反映，本身也可能带有歧视或不公，如此数据喂养的算法结果自然会烙上歧视的印记（Goodman & Flaxman，2017）。此外，算法倾向于将通过深度学习得出的结果固化或放大，一旦出现算法歧视，那么该歧视就会长存于算法之中，过去的歧视在未来的算法中还会得到巩固并加强，形成“输入—歧视—输出”的恶性循

环。在社会层面，算法歧视会放大用户的弱势（缺陷）效应，导致弱势群体越来越被无情排斥，加剧现存的不公平现象，给整个社会造成无法弥合的裂痕。

（四）算法操控

算法发展至今，无论是在商业机构还是在公共部门，基于模糊的经验或通过旷日持久的调查进行决策是符合时代潮流的（决策的科学性与准确性另当别论）。因此，算法不可避免被引入到决策当中，在某些领域，算法甚至主导了决策。当人类社会绝大部分事务离不开算法决策时，算法权力便产生了。“在一个以代码为符号的社会，权力越来越集中于算法”（Scott，2007）。算法为商业和公共机构提供了操控人类的机会，同时还刺激人类使自己的生活顺从算法的要求以避免不利决策（Balkin，2018）。借由信息传播技术的强大与便利，人类看似获得了极大的自由，实则不知不觉中“通往被算法奴役的道路”。2012年，几位研究者和Facebook的工程师合作，对70余万名不知情的用户进行了一项关乎“情绪感染”的测试，发现外部信息会对用户的情绪产生传染式影响，而用户所接受的信息是研究团队根据算法进行推荐的。用户在此研究中被当作“小白鼠”予以对待，引发了严重的伦理问题，同时也充分展示了网络平台通过算法操控用户的强大能力。同样，在2016年的美国总统大选中，剑桥分析公司（Cambridge Analytics）利用Facebook的用户信息帮助特朗普团队量身定制并精准投放广告，为其最终当选做出了重要贡献。投票给哪位总统候选人看似由选民自主选择，但算法早已通过其“无微不至”的力量影响了选民的倾向，而这种影响常常是潜移默化的，选民甚至毫无觉察。倘若信息可被强力操控，那么追求事实的真相自然变得愈发困难。

二、解释权之提出

在美国以及我国，尽管算法决策技术日益发达，且社会各界都有算法透明化的要求，然而，并没有出台针对算法规制的举措（Kaminski，2018）。而在欧洲，自上世纪90年代以来，自动化系统做出的不透明、难以挑战的决策已经引起欧盟的密切关注。尤其是随着欧盟GDPR的生效执行，解释权被不少人视为提升算法决策透明性及可归责性的关键机制。

（一）法律渊源

1995年欧盟推出《数据保护指令》（Data Protection Directive，简称DPD），该指令第12条规定数据主体有权了解其个人数据是否以及如何被处理，也有权了解数据自动化处理系统的内在逻辑，第15条规定数据主体不应成为自动化决策的对

象，以及重要的决策不能单独依据自动化数据处理系统做出（Bygrave，2001）。数据主体有权对自动化决策进行干预以便表达他们的观点，对决策进行干预意味着将数据主体纳入决策处理过程，因而被形象地称为“圈内人”（Edwards & Veale，2017）。

在DPD制定之前，《法国数据保护法》（1978）就规定数据主体在不违反版权规制的前提下，有权获得与自动化决策运行逻辑相关的信息，允许数据主体质疑决策。提供决策做出的一般逻辑与数据类型的信息是法定义务，但无须提供每个因素的精准信息，也不必揭示自动化决策系统或算法的完整代码。法国2016年出台的《数字共和国法》要求行政机关运用算法对个体进行决策时，需要对后者解释算法决策的规则。2017年，法国政府又增发一道法令（R311-3-1-2），详细要求行政机关应提供算法决策的模式、等级、处理参数、数据的来源与处理等信息（Edwards & Veale，2018）。与法国相似，英国《数据保护法案》（1998）规定数据控制者必须通知数据主体其做出决策所考虑到的因素，但无须揭露这些因素的精准权重，做出如此限制性规定的目的在于保护商业秘密。德国《联邦数据保护法》自1977年颁布以来，便占据该国数据保护体系的中心地位，同时也深受欧盟数据保护立法的影响。2002年，该法为执行欧盟DPD而得以修订，规定数据主体应被告知收集、处理或使用数据的目的，旨在保护个人隐私权，防止个人数据在使用过程中被侵害。2015年，《德国数据保护法》再次得到修订，增设数据主体不支持自动化决策时要求解释的权利，这个权利被认为是对抗自动化决策的保障机制；此外，它还允许数据主体获得自动化处理之逻辑结构的信息以及理解决策是如何形成的。

需要指出的是，由访问权所授予的数据主体获得自动化决策解释权并非新生事物。自DPD颁布以来，这个权利便已存在，并被大多数欧盟成员国的法律所采纳。DPD规定的访问权为数据主体发现数据控制者是否在处理其个人数据提供了方法；如果被处理的话，数据主体便有权获悉数据被处理的范围，从而使得数据主体审查何种数据被使用以及采取适当行动（要求修改或删除）成为可能。然而，就DPD以及欧盟一些主要成员国的立法目的而言，对计算机决策进行干预并赋予数据主体的知情权，主要在于保护信息隐私、数据安全等人身或财产权益。随着人工智能时代的到来以及算法的大规模应用，原有的法律规则已无法适应新的挑战，这促使了GDPR的制定以及“解释权”在学理层面的出现。

（二）欧盟GDPR解释权条款及内涵

2018年5月，GDPR在欧盟正式生效，这部新法律被形容为数据保护领域的

“哥白尼革命”，也被认为有效创造了解释权（Goodman & Flaxman，2017）。然而，在GDPR全文中，解释权仅被序言第71条明确提到，它规定遭受自动化决策的个体有权获得人工干预权、表达意见、获取决策如何做出的解释并有权质疑。与访问权（第15条）、数据擦除权（第17条）不同，解释权并不是GDPR明文规定的权利，而是学界根据条文的立法精神以及法律涵义所推论出的权利。学界一般认为GDPR第13-14条的告知义务、第15条的访问权以及第22条有关个人的自动化决策规定是解释权的主要来源。

首先，GDPR详列了通过数据主体（第13条）或第三方（第14条）收集、处理数据时数据控制者的告知义务。第13（2）条与第14（2）条规定，数据控制者应当向数据主体提供确保涉及数据主体的处理是公正与透明的必要信息。根据第13（2）（f）条与第14（2)（g）条，这些信息包括“存在自动化的决策，包括第22（1）与（4）条所规定的用户画像，以及在此类情形下，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”。这种义务在第22（1）条或第22（4）条中也有体现。第13-14条通知义务与第22（3）条保障措施的结合，被认为是自动化决策的事后解释权。

GDPR第15（1)（h）条的内容与第13（2)（f）条、第14(2)(g)条完全相同：数据主体有权被告知自动化决策的存在以及获得对于相关逻辑、包括此类处理对于数据主体的预期后果的有效信息。根据第15（1)（h）条，数据主体应被告知数据处理的存在、目的、逻辑以及法律后果，目的在于通过拥有这些信息，数据主体能够审视数据处理的合法性以便启动法律救济。简而言之，GDPR第13-15条赋予数据主体获得与其有关的数据处理信息以及审查数据处理合法性的权利。第13-14条指明了数据控制者的通知义务，而第15条建立了与数据主体访问权相关的权利，不同于数据控制者的通知义务，访问权由数据主体启动。

GDPR第22条规定数据主体的权利不应屈从于自动化决策,并为此提供了对抗自动化决策的保障措施。根据第22（1）条，数据主体有权反对完全依靠自动化处理（包括用户画像）对数据主体做出具有法律影响或类似严重影响的决策。第22（3）条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。

如上所述，GDPR第13-15条规定了访问权和告知权。当自动化决策发生时，第22条要求保障数据主体的权利与自由。第22条并未详尽说明具体的保障措施，但

根据第13-14条，当自动化决策（包括“用户画像”）发生时，数据主体有权获得有关自动化决策运行逻辑的有效信息，这个规定引发了两个关键问题：其含义是什么？其要求是什么——能解释算法决策吗？

三、解释权之审视

GDPR在数据保护领域所带来的革命性影响是毋庸置疑的。然而，是否如一些学者所言——GDPR创造了解释权？解释权是否是应对算法伤害的“灵丹妙药”？在为解释权欢呼之前，这两个问题值得我们认真审视。

（一）解释权是否真的存在？

针对一些学者声称GDPR包含算法决策的解释权，以牛津大学互联网研究院沃切尔教授为代表的学者对此进行了严肃质疑（Wachter，Mittelstadt & Floridi，2017）。针对自动化决策的解释可分为对系统功能的解释与对特定决策的解释两种。系统功能在此指的是自动化决策系统的逻辑、意义、预期后果以及一般功能（如系统的需求规范、决策树、预定义的模型、技术标准、分类结构等），特定决策是指做出一个特定自动化决策的原理与个人情况（如不同因素的权重、机器定义的特定案例决策机制、起辅助作用的参考信息）。在沃切尔等看来，事前解释仅用于系统功能，因为在决策做出前是不可能知道特定决策原理的，而事后解释可用于系统功能与特定决策两个方面（Wachter，Mittelstadt & Floridi，2017）。我们所讨论的解释权当为事后解释权，即当自动化决策对数据主体（相对人）产生法律上或经济上的重大影响时，数据主体（相对人）向数据控制者（算法使用人）提出异议，要求提供对具体决策进行解释，要求更正错误并进行救济的权利。

前面指出，GDPR第13-14条通知义务与第22（3）条保障措施的结合被认为是自动化决策的事后解释权，但在沃切尔等学者看来，这种说法存在明显缺陷（Wachter，Mittelstadt & Floridi，2017）。对系统功能的事前解释为13（2）（f）、14（2）（g）所明确要求。这种告知义务先于决策做出，处于数据收集阶段。由于系统功能解释在逻辑上先于决策做出，故第13-14条不能作为自动化决策事后解释权的依据（时间轴问题）。另外，第13（2)（f）条、第14（2）（g）条与第22（3）条保障措施的链接并非GDPR创造。实际上，第13（2）（f）条与第14（2)（g）条仅适用于第22（1）条与第22（4）条，而后者并没有提出保障措施以对抗自动化决策。根据第13（2）（f）条与第14（2）（g）条的条文内容，可知在自动化决策做出前，数据主体必须被提供自动化决策系统通常是如何运行的信息（比

如，目的、预测的影响等）。很显然，这并不包括有关一个特定决策如何被做出或达成的任何信息。

尽管第15（1)（h）条与第13（2）（f）条、第14（2)（g）条内容一致，但是访问权有赖于数据主体并且没有截止期限，而这并不适用第13（2)（f）条、第14（2）（g）条的时间轴（Wachter，Mittelstadt & Floridi，2017）。从第15（1）（h）条的表面意思看，数据主体在任何时候均可以要求获得信息，包括自动化决策被做出后，这使得特定决策的事后解释成为可能。然而分析第15（1）（h）条可知，“预期后果”的语义是面向未来的，表明在自动化决策发生前，数据控制者必须告知数据主体自动化决策可能的后果。如果适用于决策已被做出，那这句话就不符合逻辑。访问权并没有提到个人决策如何被做出，它更在于数据控制者提供与自动化决策有关的存在、目的与后果等信息的义务。

此外，GDPR第22（1）条规定数据主体的权利不应受自动化决策的控制，第22（3）条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。从严格意义上来讲，解释权并没有在此条文中被提及。假设自动化决策符合第22（3）（a）条（签订或履行合同）或第22（3）（c）条（明确同意）的指定条件，数据主体获得进行人工干涉，以便表达其观点和对决策进行异议的基本权利，但并不包括获得决策如何做成的解释。第22（3）条列明了合法的自动化决策所需的最低要求，在最低要求下，数据控制者并无进一步解释的义务。只要达到要求，自动化决策就是合法并符合GDPR要求的。法学界可能会继续解释“合理保障措施”的含义，并建立起包括解释权在内的对数据控制者的强制性义务，但这仅仅是一种可能的解释路径，解释权并不是当前第22（3）条强制性的法律要求。GDPR序言第71条的确提到解释权，然而序言仅为如何理解条文提供指导，本身并无法律约束力。另外，权利在被执行前应被法律明确制定，在预先没有明确义务存在的情况下，对数据控制者进行惩罚必将充满争议。依据GDPR第22条，可以得出数据主体没有被授予对特定自动化决策具有法律约束力的事后解释权。此外，根据GDPR之前的草案与谈判记录，可以看出立法者在自动化决策与用户画像上有更严格的保障，但具有法律约束力的特定决策解释权最终并没有出现，这表明立法者并不打算在GDPR中实施解释权。

综上所述，在GDPR目前的语境体系下，无论是历史沿革、立法精神还是法律条文本身的含义，证成“解释权”的存在是比较困难的。

（二）解释权是“灵丹妙药”？

证成解释权存在困难，但退一步讲，即便GDPR存在解释权，或修改和完善立法以创设解释权，算法伤害问题是否就可以迎刃而解呢？事实上，解释权并非如我们想象的那般完美，当要求算法的可解释性或透明化达到夸张的程度时，缺陷是显而易见的。

1.设计者或使用者“全能性”、算法“工具性”的打破使得解释权难以实施。在互联网发展初期，若算法决策产生不利后果，一般由设计或使用算法的网络平台承担责任。此种归责性隐含如下前提：设计者或使用者的“全能性”假设与算法的“工具性”假设——算法作为网络平台的工具，设计者或使用者有责任、有能力控制算法并对算法的运行结果具备充分的预测能力（张凌寒，2018b）。在此前提下，司法无须介入算法的内部运行，通过事后的结果审查来确定责任承担的主体。然而进入到人工智能时代，随着深度学习、神经网络等技术的发展，作为人工智能核心的算法，其功能已不再局限于按照所设计的特定程序或目的运行，发展出利用大数据并自我进化的功能。算法的准确性通常随着其复杂性的增加而增加，算法越复杂就越难解释，可解释性的要求仅在适当牺牲准确性的情况下才是可行的，然而这种情况基本上不会出现（New & Castro，2018）。算法逐渐超越“工具属性”，开始具有自身的能动性，在自动化决策体系中逐渐占据主导地位。不仅技术领域以外的人无法了解算法决策，设计者或使用者甚至也可能不再完全掌握算法的内部机制与决策过程。在“全能性”与“工具性”被打破的情况下，对算法自动化决策进行解释可能无法实现。

2.解释权行使的标准或规则并不明确。在当前的人工智能时代，算法在决策体系中逐渐占据主导地位，大至关系我们切身利益如贷款、就业、司法裁判等的决策，小至新闻推送、商品推荐等日常选择，算法自动化决策已渗入到我们生活的方方面面。在这种大环境下，若承受算法决策的相对人（数据主体）动辄行使解释权必定会使设计者或使用者陷入无穷的“释累”，姑且不论技术上是否可行，在实际中也根本无法进行。这就要求对解释权的行使设置一个合理标准。学界一般认为GDPR第22条可以作为解释权行使的标准。然而第22条是一个受限制、不清楚的法律规则。根据第22条，当自动化决策存在如下三点时，数据主体才有可能获得解释：完全依靠自动化处理；产生法律上或类似重大影响的决策；决策为签订或履行合同所必要或为法律授权或建立在明确同意基础之上。首先，影响人们生活的算法很多并不是全自动化的，实际上只是为人工决策提供支持。比如，根据2016

年ProPublica的报道，在刑事司法风险评估的算法决策系统中，种族歧视被明确排除，而且这个系统经常提供名义上的咨询（Edwards & Veale，2017）。其次，何为“决策”？除了GDPR序言第71条明确规定“可能包括一项措施”，对于何为“决策”并无其他线索。算法输出大多数是一种分类或估计，基本上具有不确定性，而这并没有约束力。另外，“产生法律上或类似重大影响”更是相当含糊。此外，关于何为“重大”，也很难存在共识。最后，只有在签订或履行合同或法律授权或明确同意时才可能获得解释权，这将使得大量在侵权法领域的算法伤害面临无法可据的尴尬境地。因此，GDPR第22条有可能使解释权形同虚设。

3.解释权之限度难以确定。算法决策的哪些方面可以解释？限度在哪里?由上文所述，对数据主体（相对人）有法律意义的解释可能分为系统功能解释与特定决策解释。这两类解释包含许多要素，那数据主体是否有权请求披露或解释全部要素呢？解释权的行使目前存在的重大限制主要在于知识产权与商业秘密之保护。相较于数据主体的解释权，各国法律更倾向于鼓励商业机构创新、保护商业秘密、促进科技发展而避免强制要求披露算法等核心代码。如前所述，法国数据保护法要求数据控制者或处理者提供决策做出的一般逻辑与数据类型的信息，但无须（至少不是全部）提供每个具体因素的精准权重，也不必揭示自动化决策系统或算法的完整代码。根据德国数据保护法，数据控制者仅需解释“决策树”的逻辑，而无须披露作出决策的特定功能的权重与参数。另外，德国SCHUFA的裁判显示数据主体没有调查自动处理系统（在案例中为信用评估）准确性的权利，因为基础公式（比如统计值、计算概率的特定元素权重、比较组等）受商业秘密的保护（Wachter，Mittelstadt & Floridi，2017）。一直以来，对于数据控制者必须披露的信息类型，尚未在欧洲各地的法院判例中得到普遍的明确范围划定。尽管商业机构的“秘密武器”（知识产权或商业秘密）不会导致数据主体得不到信息，但“秘密武器”的存在可能会使解释权“越弄越浑”。

4.即使获得解释，也可能无益于应对算法决策的挑战。尽管有助于理解算法逻辑的技术不断进步，然而以人类可理解的形式准确表达算法运行的逻辑还是众所周知的困难。以人类能理解的文字制定出来的法律，尚存在许多不能理解或争议之处，何况以复杂的数字代码形式呈现的算法。即便数据主体获得解释权，然而如何解释、解释会产生何种效果等仍掌握在商业机构或公共部门的话语体系中，普通公众甚至司法机构可能并无专业知识对解释的内容进行实质审查。若求助于专业技术人员，不仅耗费大量的时间与财力成本，而且将导致司法机构的裁判依赖于第三

方。即使算法模型、数据输入、设计权重等要素被披露，仍无法证明这个系统是被设计为带有偏见、不公正和有欺诈性的。比起明显的植入偏见的代码，大多数算法会承认无意的偏见（设计者或使用者不想被起诉，即便他们自身的道德伦理并不禁止此类行为）。另外，对用户而言，解释权的救济常常来得太晚以至于无济于事。而且，算法伤害（歧视或偏见）可能在作用于某类用户集体时才会变得明显，但这并不直接作用在个体用户上。在此种情况下，由集体抑或个体行使解释权亦是一个法律难题。这些变量可能使解释权成为一个空壳形式的权利。

只有在特定区域，解释权才确实是规制算法的有效机制。但是将解释权视为灵丹妙药或终极目标并不明智，在大多数情况下，要求透明度与可解释性将会限制创新，也无法防止潜在的危害。更有甚者，通过解释权以控制算法决策的风险创造了一个透明化陷阱，让人错以为已经有了问题的解决方案，但实际上事情并没有变得更好。

四、新路径——解释权之外

如何规制算法，阻止（至少减轻）算法决策的错误、偏见、歧视、不公对个体产生的重大影响已成为学界讨论的热门话题。解释权看似为对抗算法决策提供了有力武器，但更有可能是虚幻的景象。有学者建议在GDPR第22条增设“解释权”，并对第22条的内容进行补充细化（Wachter，Mittelstadt & Floridi，2017）。姑且不论改进的效果最终会如何，不充分利用GDPR现有规则而增设新制度并非明智之举。质言之，可在GDPR本身寻求解释权之外的解决路径，如下：

（一）诸多数据主体权之适用

GDPR赋予了数据主体诸多权利，根据其性质大致可分为两类：第一类为通用权利，包括知情权、访问权、更正权等；第二类为消极控制数据使用的权利，包括数据擦除权、数据携带权等。在通用权利中，知情权（第12条）为数据主体行使权利的前提与基础，没有知情权，数据主体便无法监督数据控制者或处理者的数据处理行为，它贯穿于数据收集与使用的全部阶段。知情权是一项立体权利，只要控制者或处理者进行了收集、处理个人数据的行为以及产生个人数据泄露的情形，均有告知数据主体相关内容的义务。访问权（第15条）规定数据主体有权从数据控制者处获得有关其个人数据是否被处理的确认结果，在个人数据被处理时，数据主体有权访问个人数据以及相关详细信息。更正权（第16条）规定数据主体有权要求数据控制者立即更正与其有关的错误的个人数据。在大数据时代，通过数据进行算法决

策成为普遍现象，而算法决策应确保所依据的数据是正确的，否则会导致错误的结果，给相对人带来不利后果。在数据收集与使用的全过程中，知情权、访问权与更正权是基本权利，它们可以针对任何个人数据的处理行为。这三种赋权的目的并不在于对个人数据使用的控制，而是避免个人数据被处理或不适当使用，从而维护数据主体的尊严。

与通用权利相比，数据擦除权、数据携带权等消极控制数据使用权利均非针对任何情形的绝对性权利，而是在区分行使条件或情形下的一种特殊控制。数据擦除权（第17条）也被称为被遗忘的权利，它规定数据主体有权要求数据控制者及时擦除其个人数据，同时阻止个人数据的进一步传播。一个需要厘清的问题是在算法决策或数据处理系统中，何种个人数据是数据主体有权处理的。数据主体对于其明确提供的数据（比如，姓名、年龄、病史记录）以用作输入算法决策系统是有权擦除的，但是否有权擦除通过分析其现实与虚拟世界中的行为踪迹而得到的观测数据？现实生活中，用户上网浏览的网页或查看的照片，GPS追踪到的位置，或其他因物理移动所产生的观测数据，目前已被网络平台大量使用。当这些观测数据使数据主体被清楚识别时，那么第4（1）条界定，当归为个人数据范畴。由于一个系统的输出可能成为另一个系统的输入，故不合格的数据应被特别擦除。在算法系统被交易或转让后，数据主体要求擦除数据，原数据控制者应当采取包括技术措施在内的合理措施告知正在处理个人数据的控制者，数据主体已经要求他们擦除与个人数据有关的链接、备份或复制。数据携带权（第20条）规定数据主体有权获得其提供给控制者的个人数据，且其获得的个人数据应当是经过整理的、普遍使用的与机器可读的，数据主体有权无障碍地将这类数据从一个控制者传输给另一个控制者。数据携带权与数据擦除权可被视为一对姊妹权利。理论上，数据主体要求网站擦除其数据的同时，也可要求将数据传输到他们自己或第三方手中，或者要求数据直接从控制者A传输到控制者B。在一些学者看来，数据携带权有效引入了竞争机制，不仅增强了数据主体对个人数据的控制，也使得算法系统对个人数据的处理不再“随心所欲”（Edwards & Veale，2017）。质言之，算法系统通过数据的收集、挖掘与使用对数据主体施加影响，进而操控其行为或损害其权益，而个人数据的可擦除或可携带使得算法系统无法有效作用于数据主体，从而减少后者遭损害的风险。

（二）数据控制者或处理者义务之适用

通过对数据控制者或处理者课以严格的义务，以保障数据主体的基本权利与自由是GDPR的基本逻辑。但GDPR的实质目的并不在于授予权利或课以义务，而是

试图创造一个减少自动化系统“毒性”的制度环境。为此，除承担一般性义务外，GDPR要求数据控制者自身参与到权利不被系统侵害的设计当中，这包括：1. 在系统实际进行中时，控制者必须实施适当的技术和组织措施以保护数据主体的权利，例如建立数据保护默认机制、匿名机制与数据最小化机制等；2. 当使用新技术处理数据可能对数据主体的权利产生高风险时，则事先必须进行数据保护影响评估；3. 公共机构、商业组织以及任何处理特殊数据的控制者必须委任数据保护官。尤其值得注意的是，GDPR所规定的数据保护影响评估（第35条）是强制性义务，它提出的“高风险”技术概念可涵盖不少算法系统。算法影响评估既可以是事前的，侧重于前瞻性分析，也可以是事后的，侧重于连续性与历史性分析。尽管存在“高风险”阈值的不确定性，但数据保护影响评估很可能成为算法系统的必需标准，特别是涉及敏感个人数据（如种族或者政治观点）的大规模处理时（Edwards & Veale，2017）。

GDPR的自愿措施对算法系统可能同样具有影响力。第40条和第42条提到为了证明数据控制者或处理者的数据处理行为符合本条例的规定，欧盟成员国应当鼓励建立数据保护行为准则与认证机制。这些规定提供了运行“大数据正当程序”权利的机会。认证可以应用于算法系统的两个主要方面：1. 通过直接指定算法的设计规范或设计过程（如所涉及的专业知识）和/或指定与输出相关的需求用以监控和评估（基于性能的标准），从而将算法认证为软件对象；2. 对使用系统进行决策的人员或过程的认证，该认证将把算法定位为在上下文中所做的使用（Edwards & Veale，2017）。在此类情形下，不仅公平或歧视等问题在认证标准中可以被考虑到，这也是一个鼓励创造更多“可解读算法”的机会。

（三）救济和责任体系之适用

GDPR赋予数据主体行政与司法两种具体的救济方式，数据主体可通过向监管机构申诉、向法院起诉的方式，寻求损害救济。若数据主体的个人数据被违法处理时，他可向其经常居住地、工作地或侵权行为地所属成员国的监管机构申诉，接受申诉的监管机构应当告知有关申诉的进展以及结果，并告知寻求司法救济的可能性（第77条）。若相关监管机构未处理申诉（或未告知申诉的进展与结果）或数据控制者（或处理者）违法处理数据侵害数据主体权利时，数据主体有权针对监管机构或数据控制者（或处理者）寻求司法救济（第78、79条）。另外，GDPR的责任体系已经相当完善，既有民事责任，也包括行政处罚。民事责任主要为获得损害赔偿，当侵权主体不止一个数据控制者或使用者，或他们同时涉及到同一处理而损害

数据主体权益时，每个数据控制者或使用者应当对损失承担连带责任，以保证数据主体能得到有效赔偿（第82条）。行政责任主要为行政罚款。行政罚款的条件一般根据违法行为、主观方面、损害结果以及减损措施等具体情况而确定，而且为保证罚款有效、适当，在罚款数额方面则根据不同的情节设置了额度标准（第83条），威慑力较大。在笔者看来，GDPR针对数据权利人和数据控制者所确立的这些救济和责任体系，基本上也都适用于算法所带来的伤害。

耶鲁大学法学院巴尔金教授曾提出“算法社会三法则”，即算法操作者是客户或终端用户的信息受托人（Information Fiduciaries）；算法操作者负有公众责任（Public Duties）；算法操作者负有不参与算法伤害的公共义务（Balkin，2017）。从上可以看出，巴尔金所提的“算法社会三法则”基本上被GDPR覆盖，而且解释权的内涵与功效也在GDPR中被分担与承接，故我们可充分利用GDPR现有的规范和救济体系，谨慎构思解释权，避免造成冗余或冲突。

五、结论和讨论

在社会的决策体系与话语体系中，算法占据主导地位并非新近才出现，但置身当下的人工智能时代，受算法影响的相对人对此并没有清晰的认识。即便算法给相对人的权益造成损害，他们也很可能没有意识到何种权益被损害。这与算法决策的不公开透明息息相关。本文梳理了算法伤害的四个层面：用户画像、算法黑箱、算法歧视、算法操纵。这四个方面基本覆盖了算法对相对人的伤害层次。面对算法的巨大能量以及应对机制的匮乏，解释权被认为是一种规制算法伤害的有效路径。本文通过分析欧盟地区的法律，证明解释权在当前的法律体系中难以证成，也非解决算法伤害的理想机制。算法决策已经渗入到社会生活的方方面面，规制算法伤害不是某个单独的权利所能包办的，而应将算法伤害问题置于整个法律体系中进行考量。GDPR作为史上最全面的数据保护法，对当今世界数据保护体系的影响是革命性的。本文认为无需刻意寻求新的规范机制，而可充分利用GDPR现有资源，最大程度解决算法伤害所带来的问题。尽管这些具体的规制措施与构想也可能存在类似于解释权的局限性，但结果衡量的标准应是现实世界的一般标准，而非无缺憾的乌托邦构想。

此外，通过对解释权进行分析，我们也可在算法解释问题上，对在我国法律语境下思考解释权可适应性问题获得启示。解释权作为一个舶来品，尤其在GDPR生效执行后，成为法学界研究的热点。但如本文所论，解释权并非规制算法伤害的

“灵丹妙药”，其作用相对有限。而且，在GDPR的体系网络中，解释权的功能基本上已经被分担与承接，强行建立解释权可能会造成冗余甚至与既有的法律体系产生冲突。我国对个人信息保护的立法是刑法先行。《刑法修正案（七）》（2009）增设“非法获取公民个人信息罪”。《关于加强网络信息保护的决定》（2012）确立了“个人信息收集使用须合法、正当、必要并经被收集者同意，而且不得出售或非法向他人提供”的基本原则。《网络安全法》（2016）则是防御、控制与惩治三位一体的网络安全保障法律，对个人信息的保护基本延续了之前确立的规则。我国关于个人信息保护的立法整体上比较粗糙，主要基于“合法、公开、必要与同意”原则。随着人工智能技术在我国的快速发展，算法伤害问题开始大量出现，面对这一局面，我国相应的法律机制明显缺位。有学者建议配置新型权利（即解释权），以弥补传统权利体系应对的不足（张凌寒，2018a）。然而，解释权不仅是一个极具争议的概念，而且也存在较大的局限性。在欧洲，对个人数据的保护被置于人权保护的高度，无论是解释权还是GDPR都是在欧洲特定的文化、民族心理以及历史背景下孕育成长起来的，它们可以给我们提供解决问题颇具价值的视角，而如果不考虑本土司法环境，照搬欧洲经验，随意规定类似于“解释权”此类的新型权利，不仅会导致水土不服，产生一批“沉睡的权利”（或僵尸权利），而且有可能起到相反作用。即便是最全面（最严格）的数据保护法律GDPR本身也存在许多争议之处，包括数据擦除权与言论自由的冲突、数据携带权与竞争法的冲突等。另外，越来越多的实证数据显示，GDPR严格的合规要求巩固了大企业的优势地位，而中小企业因在合规竞争力上难以同步被迫退出市场，其实施在给个人权利保护带来积极力量的同时，也对欧盟的技术创新以及数字经济竞争力产生显著的负面效应。在我国当前的法律语境中，在缺失配套体系的情况下，讨论建立解释权可能有失妥当。当前我们的首要任务应该是明晰欧盟个人数据保护背后的法理基础与立法精神，研究在何种范围及程度上可兹借鉴，而不是孤立地讨论解释权的可配置性问题。

（责任编辑：王思文）

参考文献 [References]

丁晓东（2017）。算法与歧视——从美国教育平权案看算法伦理与法律解释。《中外法学》，（6），1609-1623。

胡凌（2017）。人工智能的法律想象。《文化纵横》，（2），108-116。

卢克·多梅尔（2016）。《算法时代》（胡小锐，钟毅译）。北京：中信出版集团。

於兴中（2018）。算法社会与人的秉性。《中国法律评论》，（3），107-118。

张凌寒（2018a）。商业自动化决策的算法解释权研究。《法律科学》，（3），65-74。

张凌寒（2018b）。风险防范下算法的监管路径研究。《交大法学》，（4），49-62。

张潇潇（2017）。算法新闻个性化推荐的理念、意义及伦理风险。《传媒》，（6），23-25。

Balkin, J. M. (2017). The Three Laws of Robotics in the Age of Big Data. *OhioState Law Journal,* 78(5), 1217-1241.

Balkin, J. M. (2018). Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. *University of California, Davis,* 51, 1149-1210.

Bygrave, L. A. (2001). Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling. *Computer Law & Security Review,* 17, 17-24.

Edwards, L., & Veale, M. (2017). Slave To The Algorithm? Why a ‘Right to an Explanation ‘is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review,* 16, 18-84.

Edwards, L., & Veale, M. (2018). Enslaving the Algorithm: From a “Right to an Explanation “to a “Right to Better Decisions”? *IEEE Security & Privacy,* 16(3), 46-54.

Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-making and a “Right to Explanation”. *AI Magazine,* 38, 1-9.

Kaminski, M. (2018). The GDPR's Version of Algorithmic Accountability. *Technology Law,* 1-2.

New, J., & Castro, D. (2018). How Policymakers Can Foster Algorithmic Accountability. *Center for Data Innovation,* 1-40.

Perel, M., & Elkin-Koren, N. (2017). Black Box Tinkering: Beyond Disclosure In Algorithmic Enforcement. *Florida Law Review,* 69, 181-221.

Scott, L. (2007). Power after Hegemony: Cultural Studies in Mutation. *Theory, Culture & Society*，24(33), 55-78.

Tufekcl, Z. (2015). Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency. *Colorado Technology Law Journal,* 13, 203-218.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law,* 76–99.