法学研究 Chinese Journal of Law
Vol.45，No.2（2023）

算法透明层次论

安晋城 *

内容提要：算法透明的功能价值、技术基础与规范体系问题值得反思讨论。价值层面，除改进与证明两项工具价值外，算法透明还可增进算法社会的交往、理性与信任，监督算法权力，从而促进人格自由、保障人性尊严，具有内生价值。技术层面，可解释人工智能近年来飞速发展，形成了“模型为中心”的透明与“用户为中心”的透明、“内在透明”与“事后透明”、“可观察、可分解与可模拟的透明”等技术层次，为打开黑箱、实现算法透明提供了技术基础。规范层面，现行算法透明制度在规范性质上左右摇摆，应将前述价值目标、技术基础与规范体系相互耦合，建构由柔性规范、中性规范和刚性规范组成的多层次体系。

关键词：算法　透明原则　可解释人工智能　算法黑箱

算法决策的黑箱问题是当今社会的一大隐忧。〔1〕作为解决算法黑箱问题的法律方案，算法透明原则的命运颇为坎坷。理论上，学界始终没有对算法透明形成统一认识。有的认为算法透明十分重要，〔2〕有的则认为算法透明的作用微乎其微。〔3〕实务上，虽然个人信息保护法第24条明确规定，利用个人信息进行自动化决策，应当保证决策的透明度，决策对个人权益有重大影响的，个人有权要求说明，但是该规定过于笼统、抽象，一些关键问题未得到明确回答。即便《关于加强互联网信息服务算法综合治理的指导意见》（国信办发文〔2021〕7号，以下简称“算法治理指导意见”）、《互联网信息服务算法推荐管理规定》（国家互联网信息办公室、工业和信息化部、公安部、国家市场监督管理总局联合发布，以下简称“算法推荐管

* 中国政法大学民商经济法学院讲师。
本文受国家社科基金重大课题“互联网经济的法治保障”（18ZDA149）及中国政法大学创新团队支持计划（20CXTD02）资助。

〔1〕 参见马长山：《智慧社会背景下的“第四代人权”及其保障》，《中国法学》2019年第5期，第6页；Frank Pasquale, *The Black Box Society : The Secret Algorithms that Control Money and Information*, Harvard University Press, 2015, pp. 6-8。

〔2〕 参见汪庆华：《算法透明的多重维度和算法问责》，《比较法研究》2020年第6期，第165页。

〔3〕 参见沈伟伟：《算法透明原则的迷思——算法规制理论的批判》，《环球法律评论》2019年第6期，第21页。

理规定”）等部门规章制定了相对具体的规范，但这些规范在性质上自相矛盾，给实践带来了很大的困扰。

有关算法透明的讨论中，有三个问题十分引人注目：第一，算法透明有必要吗？长期以来，学者习惯于从工具的角度理解算法透明原则，在发现了替代工具后，往往怀疑算法透明原则的必要性。〔4〕第二，算法透明可行吗？机器学习算法兴起之初，其黑箱性质曾让许多学者怀疑算法透明的技术可行性。〔5〕第三，法律规范如何助力算法透明的实现？现有规范在算法透明制度的规范性质上缺乏协调安排，一定程度上阻碍了算法透明的实现。这三个问题紧密相连、环环相扣，共同决定了算法透明的理论认识与规范生成，需要一并分析讨论。

针对上述三个问题，本文尝试深入挖掘透明原则的内生价值，引入新的价值视角；系统梳理可解释人工智能技术的最新发展，引入新的技术视角；汲取现有制度资源建构逻辑清晰的规范体系，引入新的规范视角；将价值、技术与规范各自的内容层次相互耦合，形成层次鲜明的理论与规范体系。

一、从工具价值到内生价值：算法透明的价值层次

目前，算法透明通常被理解为辅助算法问责与改进算法设计的工具。一方面，在反歧视与反垄断案件中，算法透明有助于识别算法歧视与算法共谋，为执法机关的处罚提供证据；〔6〕在算法侵权案件中，算法透明有助于查明损害后果与算法决策之间的因果关系，为责任分配寻找依据。另一方面，算法透明还有助于洞悉算法系统的运作机理，提升算法模型的泛化能力与运算效率，进而优化算法系统的预测精度。〔7〕概言之，在工具价值层面，算法透明具有提供违法证据的证明价值和提升算法效能的改进价值。

（一）算法透明只有工具价值吗

算法透明一旦被理解为单纯的工具，就可能被其他工具替代，因为只要能达到同样的目的，此工具与彼工具没有实质区别。〔8〕一种否认算法透明的观点由此认为，算法透明可以被其他机制取代，没有制度化的必要。〔9〕笔者则认为，除上述工具价值外，算法透明还具有内生性价值。在统计学、经济学和社会学中，内生性（endogeneity）意指内生于系统之中而独立影响系统结果的变量属性。〔10〕笔者取其引申义，将内生于透明过程本身且相对独立地影响人的自由与尊严的价值称为内生价值。工具价值与内生价值都旨在实现人的自由与尊严，这是科技向善与“数字人权”〔11〕的基本要求，也是算法规范的主要目标。但两者仍有较大区别：工

〔4〕 参见苏宇：《算法规制的谱系》，《中国法学》2020 年第 3 期，第 172 页以下。

〔5〕 参见张凌寒：《算法权力的兴起、异化及法律规制》，《法商研究》2019 年第 4 期，第 65 页。

〔6〕 参见李成：《人工智能歧视的法律治理》，《中国法学》2021 年第 2 期，第 143 页。

〔7〕 机器学习的目的是发现隐含在数据集背后的规律，算法模型将一个数据集的规律运用到其他具有相同或相似规律的数据集上的能力，称为泛化能力（generalization ability）。See Finale Doshi-Velez & Been Kim, *Considerations for Evaluation and Generalization in Interpretable Machine Learning*, in Hugo Escalante, Isabelle Guyon *et al.* (eds.), *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Springer 2018, p. 8.

〔8〕 See Jack Balkin, *The Path of Robotics Law*, 6 California Law Review Circuit 47 (2015).

〔9〕 参见前引〔3〕，沈伟伟文，第 27 页。

〔10〕 See Aaron Hill, Scott Johnson *et al.*, *Endogeneity: A Review and Agenda for the Methodology-Practice Divide Affecting Micro and Macro Research*, 47 Journal of Management 105 (2020).

〔11〕 参见前引〔1〕，马长山文，第 8 页。

具价值重在透明所揭示的结果，内生价值则重在透明过程本身；工具价值借助算法问责等制度间接实现，内生价值则通过人类的行为心理与声誉机制相对直接地实现。一旦放弃了透明过程，内生价值将随之一并消散，保障人的自由与尊严的功能也将大大减弱，所以，内生价值因其固有性和直接性而不可替代。

内生价值主要体现在两个方面：积极促进方面，算法透明使人们在人机交互中自我调适、理性自主、进退有据，在信任算法的前提下自由发展人格；算法透明还使人们在算法正当程序中获得尊重，避免被自动化决策完全操控而丧失理性、沦为“变形人”，从而形成基于理解的信任，维持人的主体性与尊严感。消极防御方面，算法透明通过全景敞视效果监督算法权力，抵御人格自由与人性尊严面临的威胁。算法透明在积极方面直接作用于算法波及者的心理，在消极方面直接作用于算法控制者的声誉，〔12〕都可以不过分依赖其他制度而相对独立地促进或保障人格自由与人性尊严，因而具有内生价值。以下分述之。

（二）积极型内生价值：增进算法社会的交往、理性与信任

与动物的自由不同，人的自由具有更强的社会性。人类自主参加各项社会活动、参与各种社会关系，从而发展出人格自由。在社会交往中，人类努力以主体身份出现，避免沦为他人支配或操纵的对象，由此发展出人性尊严。透明在上述社会性自由与尊严的形成过程中扮演了至关重要的角色。正如学者所言，透明机制促进了交往主体间的理解与包容，改变了相互的观点和印象，调和了自身的情绪与感受，并影响了各自的决定及行动。〔13〕在算法社会中，透明同样具有上述功能。如果算法不透明，人类面对算法活动时便无法作出理性的自主选择，无法信任算法社会，最终丧失在算法社会中发展人格的机遇。算法透明则可以促进算法交往，强化理性选择，增进算法信任，助力人类在算法社会中完善自我，最终促进自由与尊严的发展。

1．交往价值

透明具有传递信息、搭建桥梁的作用，是社会交往不可或缺的前提条件。基于人脑运作不透明的生物特征，人类形成了语言、文字等透明机制，由此传递信息、沟通彼此。千百年来，人际交往之所以顺畅无阻，正是透明机制发挥了作用。〔14〕如今，人类正逐步迈向算法社会，算法系统不仅是交往的媒介，也将成为交往的对象，越来越多的生产活动、商业往来乃至日常交流都需要密切、高效的人机交互才能完成。〔15〕与人际交往类似，人机交互也依赖必要的透明机制，不畅的人机交互甚至可能酿成灾难。例如，2018 年和 2019 年波音 737MAX 机型遭遇两次空难，而引发空难的自动防失速算法系统（MCAS）根本没有在飞行员手册中出现过。〔16〕飞行员不知道该算法系统的存在，又如何实现良好的人机协作？由此可见，基于透明的人机交

〔12〕 本文所谓“算法控制者”，是指具有控制算法设计、运行和决策的权限，运用算法为社会公众提供医疗、金融投顾、个性化推荐、公共服务等服务的主体，包括大型企业和公共机构。相应地，受到算法波及影响的主体被称为“算法波及者”。

〔13〕 See Bertram Maller, *How the Mind Explains Behavior : Folk Explanations, Meaning, and Social Interaction*, MIT Press 2004, p. 63.

〔14〕 See Meg L. Jones, *The Right to a Human in the Loop : Political Constructions of Computer Automation and Personhood*, 47 Social Studies of Science 216 (2017).

〔15〕 Vgl. Bäcker in Kühling/Buchner, DS-GVO Kommentar, C. H. Beck 2020, Art. 13 Rn. 52.

〔16〕 See Andrew J. Hawkins, *Everything You Need to Know about the Boeing 737 Max Airplane Crashes*, https://www.theverge.com/2019/3/22/18275736/boeing-737-max-plane-crashes-grounded-problems-info-details-explained-reasons, last visited on 2022-12-05.

互是算法社会顺畅运转所不可或缺的一环。

从人格自由的角度看，透明是人类在社会交往中形塑人格的前提。人格的形成很大程度上依赖于其扮演的社会角色。〔17〕在社会交往中，人们会在事前综合评估自身的角色定位，事后积极吸收社会的评价反馈，自我完善、自我迭代、自我更新。与社会环境不断交互，调整适应自身的社会角色，人格方得以自由发展、逐步健全。〔18〕在算法社会中，社会角色的调适反馈器一定程度上变成了算法系统。例如，如果借款人不了解银行信贷算法系统的评价机制，事前无法判断算法对自己的定位，被拒贷后又无法获得相关评价反馈，借款人便无法与信贷算法形成交互，无法调适自身以获得发展机会，人的角色感与参与感将逐渐弱化，人格自由发展的进程便会被打断。

从人性尊严的角度看，透明是人类在社会交往中获得尊重的关键。在一方决定另一方利益的社会交往中，接受决定的一方往往不仅关注最终的结果，更关注决策的过程和理由。这种现象在司法领域尤为明显。电影《秋菊打官司》里，秋菊为讨要一个说法不屈不挠，满腔委屈希望得到倾听，满腹疑虑期待得到解释，这正是渴望正义与尊严的表现。在算法决策中，这种关注决策过程与理由的社会心理同样突出。例如，南京某律师曾起诉拼多多公司公开“砍一刀”算法，〔19〕表达了人们对自身沦为电商获客工具的深层焦虑。若算法不透明，消费者极有可能沦为电商的客体，无法在算法决策中受到尊重。对被算法波及的消费者而言，借助算法透明知悉自己被公平诚信以待，是维系主体性的重要支撑，是获得尊严感的重要源泉。

2. 理性价值

交往只是人格发展的第一步，理性才是人之为人的关键。理性不仅体现为趋利避害，更体现在理解、掌握客观规律，免遭恶意的撺掇与蛊惑，从而主宰自己的命运。理性天然地呼唤透明，透明则能开启理性之光，在透明与理性的交相辉映下，人类才能拥有真正的自由与尊严。

从人格自由的角度看，透明是人类获得理性自由的重要环境因素。正如 1983 年联邦德国宪法法院在“人口普查案”（Volkszählungsurteil）判决中所言：如果一个人无法充分、有效地掌握相关背景信息，那么他在决定做某事时，自由就会受到很大限制。〔20〕同理，复杂、陌生的算法社会充满了未知的风险，人类需要了解算法以便作出理性选择，避免算法控制者的恶意掠夺，从而实现真正的自由。如今，算法系统侵蚀人类理性的戏码正在不断上演。例如，电商平台利用算法系统分析个人的消费习惯，在超出消费者真实需求的情况下，推送个性化的商品服务链接，提供远超其清偿能力的信贷。〔21〕不了解内情的消费者无从分辨，其消费与信贷行为同理性自主的自由选择渐行渐远。

从人性尊严的角度看，透明是人类避免因蛊惑煽动而沦为他者的客体、维持理性尊严的外在条件。算法系统拥有精准把握人心而进行大规模煽动的能力，不知情者极易沦为算法控制者

〔17〕 Vgl. Niklas Luhmann, Grundrechte als Institution, Duncker & Humblot 1965, S. 65 f.

〔18〕 Vgl. Steinmüller u. a., Grundfragen des Datenschutzes, BT-Drs VI/3826, S. 86 f.

〔19〕 该案已于 2022 年 1 月 17 日一审，相关报道参见雄平观科技：《律师起诉“砍一刀”涉嫌欺诈！拼多多正式回应：页面显示数字有限》，https://finance.sina.com.cn/chanjing/gsnews/2022-01-24/doc-ikyamrmz6753793.shtml，2022 年 12 月 5 日最后访问。

〔20〕 Vgl. BVerfGE 65, 1. 这一观点在嗣后的“网络侦查案”（Rasterfahndungsurteil）等案件中得到了援引。Vgl. BVerfG Beschluss v. 5. 7. 1995, 1 BVR 2226/94.

〔21〕 See Frank Pasquale, *Humans Judged by Machines: The Rise of Artificial Intelligence in Finance, Insurance and Real Estate*, in von Braun (ed.), *Robotics, AI and Humanity*, Springer 2021, p. 121.

的客体。例如，英国脱欧公投期间，脱欧派基于算法分析向选民投放了十亿条定向广告，导致大部分选民并非出于真正的理解与意思自由而投票，沦为脱欧派的客体，丧失了民主过程的主体性。〔22〕如果算法系统更透明一些，选民在接到投票广告时能理解其中的逻辑，也许就不会被轻易地鼓动。因此，透明可以为祛除狂热、回归理性提供帮助。

3．信任价值

信任既是社会交往与理性发展的结果，反过来又可以促进社会交往与理性发展。在迈向算法社会之际，人类社会正分裂成两派：一派对算法极端恐惧，竭力抵制；另一派则无限迷信，顶礼膜拜。这两种社会心理都是在缺乏透明的背景下对算法形成的不信任或伪信任。对此，透明可以祛除恐惧与迷信，铺设信任的基石。

从人格自由的角度看，算法透明有助于消除人类对算法的恐惧，拓展算法世界的自由空间。信任建立在熟悉、了解的基础上，面对陌生而神秘的算法，人类的信任感无从建立。出于基因里对未知的恐惧，人类会逐渐拒绝进入陌生的算法世界。如果不清楚医疗诊断算法的准确率，谁敢接受它来给自己看病？如果不了解自动驾驶算法的安全性，谁敢接受它为自己代步？长此以往，数据科学技术可能趋于停滞，人类的自由空间将限于一隅。算法透明将未知的算法逐渐变为已知，人类可以借此预知前路与方向，放下对算法的戒备和恐惧，拥抱未来广阔的算法社会。因此，欲树立算法信任，算法透明是不可或缺的关键一环。

从人性尊严的角度看，算法透明建立的信任是基于理解的信任，在一定程度上具有祛魅的效果，维系人性尊严。算法迅速渗透进各个领域，展示了惊艳世人的性能与神明般的预测决策能力。在一知半解的盲目崇拜下，部分人群对算法产生了迷信。迷信的最大危害是迷信者易受造神者摆布，彻底地奉献自我，最终沦为他人的客体与手段，失去人性尊严。〔23〕人类需要警惕算法迷信，不仅要警惕算法控制者别有用心的造神行为，还要警惕人类主宰自己命运的人文主义精神逐渐枯萎。因此，算法信任应为基于理解的信任、透明的信任，而非迷信的信任。

（三）消极型内生价值：监督算法权力

算法透明还可以从消极方面监督算法权力，抵御其对人格自由与人性尊严的威胁。18 世纪中叶，哲学家边沁在与友人的书信中谈到了一种监狱设计：这种监狱由一个中央监视塔和环于四周的环形囚室构成。中央监视塔射出强光，将四周的囚室照得雪亮，无一处黑暗的角落。受强光普照的影响，即便监视塔无人值守，囚徒也会在心理上感觉有人监视自己。〔24〕这种设计被福柯称为“全景敞视主义（panopticism）”，不仅可用于看管监狱，还可用于监督权力，让掌权者在心理上感到时刻被监视，从而在行动上规规矩矩。〔25〕

透明所创造的全景敞视主义被广泛运用于监督各类权力。例如，投资世界中，投资标的信息的不对称分布使证券发行方获得了隐形权力，以公开透明为底蕴的信息披露制度因而创设。布兰戴斯大法官形象地比喻到，“阳光是最好的杀虫剂，路灯是最好的街警”。〔26〕政治关系下，秘密政治与暗箱操作强化了政府权力，公开透明的思想因而又在行政法领域扎根，形成了

〔22〕 See Marco Bastos & Dan Mercea, *Parametrizing Brexit: Mapping Twitter Political Space to Parliamentary Constituencies*, 21 Information, Communication and Society 921 (2018).

〔23〕 参见［德］韩炳哲：《透明社会》，吴琼译，中信出版社 2019 年版，第 25 页。

〔24〕 See Jeremy Bentham, *The Panopticon Writings*, Verso 1995, p. 35 f.

〔25〕 参见［法］米歇尔·福柯：《规训与惩罚》，刘北成、杨远婴译，三联书店 1999 年版，第 219 页以下。

〔26〕 See Louis Brandeis, *Other People's Money and How Bankers Use it*, Frederick A. Stokes Company, 1914, p. 92.

开放政府的理念与制度，〔27〕推动了以美国 1966 年信息自由法为代表的政府信息公开浪潮。政府有义务开放其政务资料供公众查阅复制，说明行政行为的理由，让相对人对相关行政程序有所了解、必要时提出质疑。信息披露与开放政府即全景敞视主义的体现。

如今，算法决策在商业与行政领域日渐普及。普通人的日常生活受到了算法评价、排序与推荐的深刻影响，人类逐渐进入算法主宰的世界。在尚未充分规制的算法世界，算法即律法、算力即权力，算法控制者悄然掌握了分配社会资源的隐形权力，蜕变为人格自由与人性尊严的主要威胁。将算法系统与算法控制者置于全景敞视之下，是保障数字人权的重要机制。虽然冷冰冰的算法系统不受心理因素制约，但算法背后的控制者却有血有肉，他们是算法权力的真正拥有者。如果其主导的算法活动得以全景敞视，算法控制者将因心理压力和声誉机制而自我监督。2018 年前后，脸书公司的算法系统侵害隐私问题屡屡曝光。出于对用户情绪和社会关切的考虑，该公司陆续关闭了“面部识别”“附近的朋友”等敏感算法系统，删除了相关的数据及个人信息。〔28〕显然，透明过程所激发的用户情绪与社会关切是脸书公司自我约束的主要动因，这充分显示了透明机制在规训算法权力方面的重要作用。

全景敞视的监督作用与为算法问责提供证据的监督作用不同。后者是算法透明工具价值的体现，规制重点在于归责与救济，依赖问责机制发挥监督作用。受限于严格的法律要件，算法问责在很多情况下难以实现。前者是内生价值的体现，规制重点在于预防和监督，直接作用于掌权者的心理与声誉而独立发挥监督作用。即便算法问责搁浅，全景敞视也可以发挥监督作用。这些差别在“美团配送时间预测算法事件”中表现得尤为明显。据媒体报道，美团配送时间算法压榨骑手，危害交通安全。若针对该算法开展问责，在构成要件符合性上其实存在较大难度。事实上，该事件最终通过透明机制得以处理：美团为维护公众形象主动披露并改进了相关算法。〔29〕可见，全景敞视不依赖问责机制而独立发挥监督作用，无法被算法问责取代，这就是算法透明消极防御型内生价值之所在。

（四）小结：价值层次的规范启示

由上述论证可知，因内生价值的存在，算法透明不再可有可无，而是算法治理体系中不可或缺的一环。即便摆脱了工具价值，算法透明也可以得到正当性证成。这对于否定算法透明的工具论给予了有力的回应。

更重要的是，工具价值与内生价值的区分启发了下文规范性质的配置思路。工具价值可被替代，旨在实现工具价值的制度便不宜被赋予过强的约束力。反之，内生价值不可替代，旨在实现内生价值的制度则应被赋予更强的约束力。本文第三部分将展开更详细的论述。

二、从算法黑箱到算法可释：算法透明的技术层次

由于人工智能算法涉及的数据量超出了人类的运算能力，对数据的特征提取具有很强的随机性，加之部分机器学习算法模型的数据处理过程隐蔽，人工智能算法确实呈现了黑箱特征。

〔27〕 See Wallace Parks, *The Open Government Principle: Applying the Right to Know Under the Constitution*, 26 The George Washington Law Review 1 (1957).

〔28〕 See Lauren Sailing, Daniel Cohen *et al.*, *Not Close Enough for Comfort: Facebook Users Eschew High Intimacy Negative Disclosures*, 142 Personality and Individual Differences 103 (2019).

〔29〕 参见《外卖骑手，困在系统里》，https://zhuanlan.zhihu.com/p/225120404，2022 年 12 月 5 日最后访问。

另一种否定算法透明的论点便由此认为，算法透明在技术上无法实现，实乃空中楼阁。〔30〕

（一）算法黑箱无法打开吗

从当前技术发展水平看，算法黑箱并非不可打开。近年来，可解释人工智能技术（explainable artificial intelligence，XAI）突飞猛进，算法透明的技术障碍逐渐被攻克。〔31〕例如，强化学习算法时时处于迭代更新的过程中，此刻得到解释的算法在迭代后便不复透明，给算法透明带来巨大困扰。对此，强化学习可解释性的研究者通过行为克隆、逆强化学习、策略分解等方法，解释强化学习中环境状态转移的内部规律、任务目标与过程状态序列之间的关联，从而预测算法系统依据环境变化所作出的决策，〔32〕破解算法不断更新给算法透明制造的困局。可见，算法黑箱论存在历史局限性，有必要在技术层次上对算法透明进行系统梳理。

从算法透明的目的来看，有些透明方法重在获取与算法模型密切相关的参数设置、特征权重、运行逻辑等宏观全局信息，由此获得的透明被称作“模型为中心”（model-centered）的透明；有些透明方法则重在获取与用户密切相关的输入数据、决策依据、因果关联等微观局部信息，由此获得的透明被称作“用户为中心”（subject-centered）的透明。〔33〕因这种分类对算法规范的设计具有重要影响，本文以此作为算法透明技术层次的界分标准。

（二）模型为中心的透明

从算法宏观治理的角度而言，算法模型的整体运行逻辑是人类迈进算法时代的必修课，是人类参与算法交往、保持理性自主、构建算法信任的阶梯。有些模型本就具有内在透明性，部分维度不透明的模型也可通过解释技术变得透明。因此，了解、掌握算法模型的运行规律并非天方夜谭。

1．内在透明性

某些算法模型本身就有一定程度的透明性，本文称为内在透明性。例如，音乐 APP 可以根据用户访问数据预测未来音乐流行趋势，其使用的线性回归模型本质上是在给定函数类型和约束条件的情况下求解自变量与因变量之间的回归系数。其基本思想和求解方法——“最小二乘法”为我国高中数学必修课的内容，〔34〕可见其透明程度之高。内在透明的算法模型还包括逻辑回归模型、决策树模型、K 近邻模型、规则学习模型、广义加性模型以及贝叶斯模型等。〔35〕这些模型的内在透明性包括三个方面：

其一，可观察性。如果一个算法模型的运行逻辑可以被人类从外界轻易观察，任何环节的错误也可以被轻易地检验，这个模型就具有可观察性。可观察性强调从整体视角出发、以人类熟悉的形式感知与理解模型运行逻辑，上述线性及逻辑回归模型便具有这一特征。一些智能推荐系统应用的 K 近邻模型也是如此：在一个训练数据集中，找到与一个新的输入实例最邻近

〔30〕 参见陈景辉：《算法的法律性质：言论、商业秘密还是正当程序？》，《比较法研究》2020 年第 2 期，第 131 页。

〔31〕 See David Gunning, *DARPA's Explainable Artificial Intelligence Program*, 40 AI Magazine 44 (2019).

〔32〕 See George Vouros, *Explainable Deep Reinforcement Learning: State of the Art and Challenges*, 55 ACM Computing Surveys 1 (2022).

〔33〕 See Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a Right to an Explanation is Probably not the Remedy You are Looking for*, 16 Duke Law & Technology Review 18 (2017).

〔34〕 参见人教社课程教材研究所编著：《普通高中教科书 数学 选择性必修 第三册》，人民教育出版社 2020 年版，第 110 页以下。

〔35〕 See Alejandro Arrieta, Natalia Díaz-Rodríguez *et al.*, *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges towards Responsible AI*, 58 Information Fusion 82 (2020).

的 K 个实例。如果这 K 个实例大多归属于某个类别，该输入实例就可以分类到该类别之中。这一思路取自人类对事物分类的思维方式，具有较高的可观察性。当然，这里的“轻易”是相对而言的，锚定人类的平均认知水平。相反，黑箱模型无法被人类从外界轻易地观察、理解与检验。例如，深度神经网络运用多层隐藏层处理数据，这段数据处理过程无法被人类从外界轻易、直观地观察、理解与检验，〔36〕只能通过后文介绍的事后透明方法模拟与近似。

其二，可分解性。如果一个算法模型的各个部分，包括数据输入、参数设置、计算流程等可以被良好地阐释说明，这个模型就具有可分解性。可分解性强调对模型组件的阐释和说明。可观察的模型未必可分解。未贴标签数据的无监督学习常用的 K 均值聚类模型虽然与上述 K 近邻模型的基本原理如出一辙，具有可观察性，但它需要随机选取 K 个中心点，计算 N 个样本点与 K 个中心点的欧氏距离，并对这一过程迭代，直至中心点收敛。〔37〕这种模型虽然原理清晰，但参数设置是随机的，无法被良好地阐释说明，不具有可分解性。

其三，可模拟性。如果一个算法模型可以由人类模拟，则其具有可模拟性。当涉及的参数过于复杂或输入的数据量过于庞大时，可观察或可分解的模型可能丧失可模拟性。可模拟性强调对模型整体规模与复杂度的控制。例如，简单决策树模型具有良好的可观察与可分解的性质，但是泛化能力与预测精度不高。为加以改进，需要将多个决策树聚合迭代，急剧提高模型的数量与复杂度，导致超出人类可模拟的极限。相反，神经网络的鼻祖与基本单元——感知机却因计算量小而具有极强的可模拟性。

从上文梳理可知，许多算法模型本身具有内在透明性。这些内在透明模型的应用范围远远广于否定论者津津乐道的黑箱模型，从技术层面否认算法透明可行性的论点值得商榷。

2．事后透明性

不可否认，部分机器学习算法模型因参数维度高、输入数据规模大、部分流程隐蔽、特征提取随机而缺乏部分维度的内在透明性，如梯度提升决策树、随机森林、支撑向量机、深度神经网络等。为此，计算机科学界研发了多种方法对这些模型进行事后解释，〔38〕以补全其欠缺的透明维度。以卷积神经网络为例，它使用多层隐蔽的卷积层提高运算能力，由此产生了黑箱属性。不过，最后一层卷积层具有反馈信息的特质和丰富的语义结构。研究者依靠上述属性提出了积分梯度归因、反卷积、类激活映射等方法，利用最后一层卷积层的语义结构生成特征图，呈现模型的特征选取机理，〔39〕弥补了可观察性与可分解性。

再如，模型无关局部解释方法先选取算法模型中复杂度低且对模型整体有重要影响的部分，以线性回归等简单模型对该部分建模，求得一个与黑箱模型近似的结果，再结合部分依赖图、个体条件期望图、累积局部效应图等方法，分析部分特征如何影响全局，进而实现全局解释，弥补了可模拟性。

这充分说明，黑箱模型也具有透明的可能性。与内在透明模型相比，黑箱模型的透明化需

〔36〕 See Riccardo Guidotti, Anna Monreale *et al.*, *A Survey of Methods for Explaining Black Box Models*, 51 ACM Computing Surveys 93 (2018).

〔37〕 See Mark Ming-Tso Chiang & Boris Mirkin, *Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads*, 27 Journal of Classification 3 (2010).

〔38〕 See Mengnan Du, Ninghao Liu *et al.*, *Techniques for Interpretable Machine Learning*, 63 Communications of the ACM 68 (2020).

〔39〕 See Matthew Zeiler & Robert Fergus, *Visualizing and Understanding Convolutional Networks*, in David Fleet, Thomas Pajdla *et al.* (eds.), *European Conference on Computer Vision*, Springer 2014, pp. 818-833.

要更多的技术投入，将给算法控制者带来额外的成本。

（三）用户为中心的透明

从用户微观权益的角度而言，算法模型的整体逻辑不是十分重要，用户往往更关注与其密切相关的内容，并且十分看重解释的直观性。〔40〕由此，计算机科学界以用户为中心提出了其他透明方法。

以个性化推荐系统所采用的协同过滤算法为例，它收集用户的行为数据，基于聚类、关联规则挖掘、矩阵分解等模型拟合用户对物品的评分，再针对“用户—物品评分矩阵”计算皮尔森相关系数、余弦相似度等指标，从而作出推荐。其中，矩阵分解模型利用隐因子拟合用户评分。由于隐因子在现实中缺少实际含义，因而矩阵分解的结构难以解释。对此，学界尝试在算法推荐系统中引入对用户评论的情感分析或者用户与物品的交互行为等显式特征进行约束，将矩阵分解的隐因子与显式特征对齐，从而将隐因子模型变为显因子模型，继而以文本对话或者词云的形式将上述推荐规则表达出来。〔41〕这种透明方法紧密围绕用户的理解需求展开，不再关注算法模型的基本逻辑、参数权重及复杂度控制，具有用户为中心的特征。

再如，反事实解释更改原始的输入数据，以观察结果的相应变化，也属于用户为中心的透明方法。〔42〕例如，微幅改变信用卡申请者的特征输入数据，观察他们的申请是否会被算法系统重新接受，从而判断被改变的特征是否为算法系统拒绝申请的原因。沙普利加性解释工具即可实现上述功能。它通过添加新的特征或改变现有特征的权重，测试所产生的扰动对最终结果的影响，在此基础上为输入变量的特征相关性打分，以反映算法模型的决策标准。派森（Python）官方已将它开源提供给所有编程人员使用。这种解释符合人们希望了解此事发生而彼事未发生的心理需求，在满足用户微观权益方面具有重要意义。

用户为中心的透明无法提供算法模型的详尽信息，但给用户亲近算法、理性地信任算法创造了契机，对于实现算法透明的内生价值具有重要意义。

（四）小结：技术层次的规范启示

上述关于算法透明技术层次的讨论可带来三方面启示：

其一，算法透明在技术上无法实现的论点显然站不住脚。有些模型本身就具有良好的内在透明性，有些模型虽然不具有内在透明性，但可以通过各种各样的事后透明方法得以弥补。

其二，算法透明具有不同的层次与实现方法。有的以模型为中心，有的则以用户为中心；有的算法模型具有内在透明性，有的则需采用事后解释方法；有的模型具有分解式透明，有的则具有模拟式或观察式透明。因不同的算法透明技术所能提供的信息种类不同，算法透明制度不宜采用一刀切的规制思路。

其三，在算法设计上，内在透明模型无需进一步配套研发事后解释方法，成本更低、难度更小。在算法产业界，已有专家提出尽可能使用透明度更高的模型架构算法，在设计中实现算法透明，即“经由设计的透明”。〔43〕该制度的定位将于后文详细讨论。

---

〔40〕 See Andrew Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 Fordham Law Review 1085 (2018).

〔41〕 See Yongfeng Zhang *et al.*, *Explicit Factor Models for Explainable Recommendation based on Phrase-Level Sentiment Analysis*, in ACM SIGIR Conference on Research & Development in Information Retrieval 2014, pp. 83-92.

〔42〕 See Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*, 31 (2) Harvard Journal of Law & Technology 841 (2018).

〔43〕 See Thomas Wischmeyer, *Artificial Intelligence and Transparency: Opening the Black Box*, in Thomas Wischmeyer (ed.), *Regulating Artificial Intelligence*, Springer 2020, p. 75.

三、从柔性规范到刚性规范：算法透明的规范层次

针对算法透明缺失的问题，法学界提出了诸多制度，如算法解释权、算法参数备案、算法开源、算法影响评估、算法合规审计、算法决策告知、经由设计的透明等。这些制度的概念、意义与原理已有诸多讨论，但如何配置不同制度的强制力仍缺乏研究。国家互联网信息办公室2021年9月牵头制定的“算法治理指导意见”要求算法“应”透明，“算法推荐管理规定”的征求意见稿也遵从了类似的强制性方案，但2021年12月发布的“算法推荐管理规定”的正式稿则仅仅“鼓励”算法透明，说明立法者在算法透明规范的强制力上抱有疑虑。

（一）价值、技术与规范的耦合

由于规范性质的错置，现有透明制度未产生预期效果。有的规范本应得到严格遵守，却因缺乏严厉的法律效果而被虚置，如算法决策告知；有的规范未尊重具体的利益情况与技术情况，遭到了实务界与理论界的强烈反对，如算法解释权；〔44〕有的规范所追求的目标是正当的，但普遍遵守则有些强人所难，如算法开源。只有为不同的制度分门别类地配置规范性质，适应被规范主体对规范自由度的不同需求，才能产生令人满意的规范效果。

就被规范主体的自由度而言，一种传统的分类是硬法与软法。硬法具有较强的稳定性和强制力，主要由立法主体自上而下制定，基本表现形式是法律与行政法规；软法具有较强的适应性和灵活度，主要由被规范主体自下而上形成，基本表现形式是行为准则、指南和最佳实践。这种分类适应了算法多元治理的基本思路，足资借鉴。但是，有些规范可能既不宜制定成毫无回旋余地的硬法，也不宜制定成约束力过弱的软法。这种二元对立的规范划分，不适应算法治理实践。域外法理学近来提出，硬法与软法之间还存在中间地带。〔45〕受此启发，笔者将硬法与软法的二分发展为刚性规范、中性规范与柔性规范的三分。其中，刚性规范属于传统的硬法，往往通过使合同无效、施加行政或刑事处罚以及惩罚性赔偿保障其强制力。中性规范本质上也属于硬法，但被规范主体在严格遵守法定程序的前提下，可以根据具体情况灵活决定是否履行相关实体性义务，因而比刚性规范具有更强的适应性。柔性规范则属于传统的软法，但当代法理学普遍主张以“遵守或解释”〔46〕机制加强软法的约束力，本文的柔性规范就是此类强化型软法规范。

在算法治理领域，规范强制力的配置是有章可循的。结合上文所述，价值是规范所追求的目标，技术是规范所依赖的前提。价值与技术相互耦合，共同决定了规范的性质。

透明价值决定了柔性规范与其他规范的划分。实践中，有些算法模型公开后将侵害商业秘密与知识产权，或者给不法分子留下可乘之机。〔47〕是故，立法者不仅要实现透明价值，还要处理透明价值与其他价值的冲突。在处理价值冲突时，工具价值层次的透明需要作出更多让

〔44〕 See Anrew Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 International Data Privacy Law 233 (2017).

〔45〕 See Rolf Weber, *Artificial Intelligence ante portas: Reactions of Law*, 4 Multidisciplinary Scientific Journal 486 (2021).

〔46〕 “遵守或解释”机制最早伴随着英国1992年针对公司治理颁布的《吉百利准则》(Cadbury Code)出现，后来被欧盟《特殊公司年度审计指令》(Directive 2006/46/EC on the annual accounts of certain types of companies)等多部规范继受，如今已成为法理学普遍承认的一项立法技术。See John Lowry & Arad Reisberg, *Pettet's Company Law*, 4th ed., Pearson 2012, p. 207.

〔47〕 See Mike Ananny & Kate Crawford, *Seeing without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability*, 20 New Media & Society 973 (2018).

步，而内生价值层次的透明不应作出过多让步。由此，可以将以实现工具价值为主的制度配置为柔性规范，将以实现内生价值为主的制度配置为刚性或中性规范。

算法技术进一步决定了中性规范与刚性规范的划分。在实现内生价值的制度中，有些制度依赖算法解释技术，不能在实体上罔顾技术细节而设置一刀切的规定，只能由熟稔技术的算法控制者在严格遵循法定程序的前提下，结合技术细节自行决定透明与否以及相应的层次和内容；有些则无需借助解释技术，通过强化法律责任就能获得良好的规制效果。前者可以配置为中性规范，后者则可以配置为刚性规范。

（二）柔性规范

柔性规范主要实现算法透明的工具价值，通常表现为基于公众讨论和行业共识而形成的治理准则或行为准则，体现了行业最佳实践，为被规范主体提供了改善治理的样板，如各式各样的“人工智能治理准则”。传统理论中，是否遵守柔性规范完全由被规范主体自行决定。但是，公司治理领域发展出了“遵守或解释”机制，即被规范主体可以不遵守该规范，但应解释不遵守的理由，从而使柔性规范具有一定的软约束力，这值得算法治理领域借鉴。

1．经由设计的透明

“算法推荐管理规定”第 12 条提出：“鼓励算法推荐服务提供者优化检索、排序、选择、推送、展示等规则的透明度。”算法控制者利用算法设计提升透明度，即上文所言的“经由设计的透明”。结合前文的技术层次讨论，该规定具有两方面含义：一方面，算法设计时应首选具有内在透明性的模型。例如，在上文提到的协同过滤算法中，矩阵分解的黑箱模型可以被关联规则挖掘的透明模型替代，由后者集成的个性化推荐系统更具透明性。另一方面，即便选择黑箱算法，也应采取必要且适当的技术工具改进算法模型，优化模型的透明度。例如，利用谷歌公司研发的模型卡片机制、语言诠释性工具与张量板工具，可以提升黑箱模型的透明度。〔48〕不过，选择何种模型以及是否改进模型毕竟是算法控制者的自由。根据技术中立原则，法律不宜强行扶持某一技术，打压其他技术。两相权衡，技术自由中立发展的价值压倒了改进算法的工具价值。因此，经由设计的透明只能设置为柔性规范。另外，高透明度算法具有亲人类性和可解释性，在节省算力和提升性能方面具有独到的竞争优势，有助于提升企业竞争力，因而改进价值可以通过技术竞争间接实现。为了避免竞争机制失效，立法者可以采用“遵守或解释”机制，增强“经由设计的透明”的约束力。

2．算法开源

备受争议的向社会公开算法源代码的制度，〔49〕也可以设置为柔性规范。由于普通民众几乎无法阅读、理解源代码，算法开源在实现人格自由与人性尊严的内生价值方面效果甚微。但是，算法开源后，算法的歧视与缺陷可以由法律、技术专业人士检验代码获知，因此，算法开源可以实现证明价值，提供一种“鱼缸式透明”。〔50〕网络世界存在众多开源社区，大型科技

〔48〕 See Google AI Blog, *Introducing the Model Card Toolkit*, https://ai.googleblog.com/2020/07/introducing-model-card-toolkit-for.html, last visited on 2022-12-05.

〔49〕 2017 年，美国纽约市尝试要求利用算法系统实施处罚的行政机构在其官网公布其系统的源代码，但遭到了比较大的反对，最终搁浅。

〔50〕 鱼缸式透明（fishbowl transparency），指的是公众只能了解信息披露主体正在做什么，与之相对，说理式透明（reasoned transparency）额外要求信息披露主体说明这样做的理由。See Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 Administrative Law Review 1 (2019). 与中性规范和刚性规范不同，算法开源制度只能让公众了解算法控制者在做什么，提供一种较为初级的透明。

公司的许多项目及其算法组件的代码都是开源的。〔51〕开源社区、开源项目和开源代码会形成一个竞争性的算法代码市场，开源算法因其透明和可信任感将在竞争中获得优势。因此，自由市场的力量会推动算法开源，国家不必强制。更重要的是，公布算法源代码涉及算法著作权、商业秘密，而且可能为戏弄算法打开方便之门，国家也不宜强制。〔52〕两相权衡，证明价值只能退居其次。综上所述，可以在现有的开源实践、公众讨论和行业共识基础上制定算法开源准则，鼓励算法控制者尽量开源。若选择不开源，则应作出合理解释。

（三）中性规范

当某一制度的运行涉及复杂的条件、无法简单地一刀切规定时，该制度适合设置成中性规范。例如，公司法第 33 条对会计账簿查阅权的规定就属于中性规范。公司可以按照法定程序审查查阅目的是否正当，并在 15 日内予以答复，拒绝查阅的应说明正当理由。双方对查阅有争议的，可以提交法院审查处理。上述规范中，程序内容具有刚性，当事人必须遵守特定程序，在特定时间内完成特定行为；实体内容则具有柔性，当事人可以结合具体场景与情事，判断是否披露以及披露的内容、方式与细节。算法解释权、参数报备等透明制度与之类似，涉及内容各异的算法技术，难以一刀切地规定，应强化程序而柔化实体，给被规范主体更多的自由选择空间。

1．算法解释权

算法解释权与股东知情权较为相似，宜设置为中性规范。一方面，算法解释权主要用来实现内生价值，需要更强的约束力；另一方面，不同算法模型的技术细节差异巨大，无法设置统一的解释事项。对此，程序刚性、〔53〕实体柔性的中性规范可以满足上述需求，而被视为算法解释权法律基础的个人信息保护法第 24 条过于笼统，需要司法解释或部门规章予以细化。具体而言，相关规范应列出潜在的解释项目，如算法控制者的主体信息、算法的影响、算法模型信息等。其中，算法主体信息和算法的影响为不得拒绝的解释项目，而涉及技术细节的算法模型信息为可以拒绝的解释项目。解释权人须在书面请求中具体指出解释内容。涉及算法模型的，算法控制者收到请求后可以根据模型的技术特点决定是否解释，并在特定时间内答复，拒绝解释的应当说明正当理由。如果解释权行使产生了争议，双方可以诉诸法院解决。

判断某项内容应否解释时，“模型为中心”与“用户为中心”的区分可以发挥一定的辅助作用。算法解释权原则上不侧重于向用户详细讲解系统的运作细节或整体逻辑，而是侧重于提供与用户个体密切相关的内容。〔54〕控制者如遇到用户要求解释算法的全局运行，原则上可以基于技术成本的原因拒绝。当然，算法控制者愿意主动解释说明的，亦无不可。〔55〕反之，对于仅涉及个体的局部解释请求，如用户询问“我为什么获得了这个决策结果”“我该如何改变以获得不同的决策结果”“与我相似的用户的基本画像为何”“我的情况是否比较特殊，以至

〔51〕 例如，美团公司便开源了诸多算法，并供公众自由下载。参见 https://tech.meituan.com，2022 年 12 月 5 日最后访问。

〔52〕 参见许可：《驯服算法：算法治理的历史展开与当代体系》，《华东政法大学学报》2022 年第 1 期，第 112 页。

〔53〕 关于算法解释权应设置为程序性规范的观点，参见丁晓东：《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》，《中国法学》2022 年第 1 期，第 113 页以下。

〔54〕 See Margot Kaminski, *The Right to Explanation, Explained*, 34 Berkeley Technology Law Journal 189 (2019).

〔55〕 例如，字节跳动公司曾在字节跳动微信公众号上委托其算法架构师向公众说明今日头条推荐算法的运行逻辑。参见《今日头条推荐系统原理》，https://cloud.tencent.com/developer/article/1052655，2022 年 12 月 5 日最后访问。

于目前的算法系统无法准确决策”，〔56〕控制者原则上不得拒绝解释，而应通过反事实解释的方法予以答复。

2．算法参数备案、影响评估与合规审计

算法参数备案、影响评估与合规审计要求算法控制者将算法的原理、意图和影响向监管机关、专业机构、审计部门乃至社会公众适当公开。〔57〕它们同样以实现内生价值为主要目的，涉及不同模型的技术细节，也宜设置为中性规范。

一方面，算法模型的可透明程度、透明方法各不相同，统一的刚性规定确实有些强人所难。比如，上文提及的 K 均值聚类模型不具备可分解的透明维度，其初始参数 K 的取值逻辑无法详细地描述与备案。况且，算法技术在不断发展，立法者无法预言未来哪些内容可以透明，因而只能柔化实体内容，将之交由算法控制者、评估者及审计者灵活决定。另一方面，与柔性规范相比，参数备案、影响评估和合规审计又应有一定的约束力。美国的算法影响评估便因为强制性制裁措施不足，没有取得应有的效果。〔58〕因此，应以行政处罚等方式强化上述制度的程序要求。

虽然“算法推荐管理规定”第 24 条、个人信息保护法第 54 条、第 55 条对上述规范的性质语焉不详，但这些制度在实践中都体现出了中性规范的特征。关于算法参数备案，笔者注册、登录算法备案系统后发现，〔59〕除算法主体信息和算法基本信息为必填项目外，算法模型、算法训练数据集及来源等算法详细属性信息均明确注明“选填”，不填写不影响备案。可见，算法备案系统的设计者考虑了算法模型的技术特点，没有强迫要求备案全部算法信息。关于算法影响评估与审计，这两项制度都需要服从算法的技术特性。算法影响评估需要借助测试数据集，从算法的残差、拟合情况、鲁棒性、〔60〕召回率、准确率、覆盖率与更新率等技术方面开展。不同的算法模型，选用的评估指标不同，算法影响评估所需的算法信息也不同。〔61〕算法审计也遵循这一原理。〔62〕我国的算法影响评估与合规审计尚无详细的规则，上述立法思路可资借鉴。

判断应否提供某项信息时，“模型为中心”与“用户为中心”的区分同样可以发挥一定的辅助作用。这三个制度显然重在模型为中心的透明，具体涉及建模意图、模型种类、模型逻辑、初始参数、数据描述、训练过程及系统评价等算法系统的核心内容。在报备、评估、审计时，上述信息是否均须提供，可以由算法控制者、评估者与审计者依据严格的程序、结合算法模型的技术特点自行判断：涉及内在透明模型的，可以按照“可观察、可分解与可模拟的透明”的维度，要求相关信息透明化；涉及事后透明模型的，可以根据可解释人工智能技术的客观发展水平，要求相关信息透明化。

---

〔56〕 参见前引〔33〕，Edwards 等文，第 18 页。

〔57〕 See Gregory Falco, Ben Schneiderman *et al.*, *Governing AI Safety through Independent Audits*, 3 Nature Machine Intelligence 566 (2021).

〔58〕 See Brent Mittelstadt, *Principles Alone Cannot Guarantee Ethical AI*, 1 Nature Machine Intelligence 501 (2019).

〔59〕 参见互联网信息服务算法备案系统，https://beian.cac.gov.cn/#/index，2022 年 12 月 5 日最后访问。

〔60〕 鲁棒性（robustness），是指系统抵御异常情况干扰的能力。参见赵长安、贺风华编著：《多变量鲁棒控制系统》，哈尔滨工业大学出版社 2011 年版，第 4 页。

〔61〕 参见张欣：《算法影响评估制度的构建机理与中国方案》，《法商研究》2021 年第 2 期，第 112 页。

〔62〕 See Bryan Casey, Ashkan Farbangi *et al.*, *Rethinking Explainable Machines: The GDPR's Right to Explanation Debate and the Rise of Algorithmic Audits in Enterprise*, 34 Berkley Technology Law Journal 143 (2019).

（四）刚性规范：算法决策告知义务

算法决策告知义务得到了“算法推荐管理规定”第16条等规范的肯认。据此，算法控制者必须显著标识算法的部署情况，充分提示算法的潜在风险。该义务旨在保障算法波及者知晓算法决策的存在，从而自由地选择算法决策或人工决策，实现上文阐述的交往、理性、信任与监督的内生价值。同时，该义务不涉及复杂的技术细节，不惧一刀切的规定，反而需要严格的规定强化其约束力。因此，算法决策告知适合设置成刚性规范。

刚性规范的约束力主要体现在法律效果上，要么伴随着严厉的法律责任，要么直接影响相关法律行为的效力。遗憾的是，除数额不高的行政处罚（“算法推荐管理规定”第31条）外，违反算法决策告知义务目前缺乏严厉的后果。为加强算法决策告知义务的约束力，应运用行政法的行政行为撤销权与民商法的产品责任，区分公私法领域分别补充建构相应的法律效果。

在行政机关未履行告知义务的情况下，行政相对人丧失了选择其他决策方式的机会，可能错失更好的决策结果。对此，单纯的国家赔偿于事无补，只有让相关行政行为失去效力，才能救济行政相对人。加之算法决策本质上属于具体行政行为，具有可撤销性，因此，不告知便作出算法决策的，行政机关应被认定为滥用职权、违反正当程序，基于该决策的行政行为因存在重大程序瑕疵而可被撤销。例如，行政机关未明确告知而使用行政罚款算法的，相对人可以在法定期间内依程序申请处罚机关、上级机关及对应的法院撤销相关处罚决定。处罚机关和上级机关在行政监督过程中，也可以主动撤销相关决定。撤销后，该决定自始无效，相对人可以要求行政机关人工裁量，或者采用透明算法而非黑箱算法，以避免算法评估的错误或偏见。这种主张不仅是对“技术性正当程序”〔63〕的回应与补强，而且现行行政法规范即可提供法律依据，立法成本极低。

在市场主体未履行告知义务的情况下，撤销权往往无法充分满足用户的权利诉求。例如，针对特斯拉汽车近几年频发的刹车失灵事故，业内较有说服力的观点认为，此类事故大概率与特斯拉的“单踏板模式”有关，即鼓励司机仅使用电门踏板，减速和刹车由算法系统通过检测司机松开电门踏板的力度和幅度自动完成，这种改变司机驾驶习惯的行为在紧急时刻却可能酿成大祸。在特斯拉的宣传中，单踏板模式的巨大风险始终没有被充分揭示，且大部分车主不知如何关闭该模式。〔64〕在上述案例中，特斯拉履行算法决策告知义务明显存在瑕疵，属于广义的未履行告知义务，且事故已经酿成损害，撤销无济于事，宜运用产品责任与消费者保护的基本法理。市场主体的算法决策构成其产品和服务的一部分，根据消费者知情权的要求应明确告知用户算法决策的存在，让消费者洞悉算法系统的潜在风险，从而作出理性选择，否则，可以解释为警示瑕疵，适用产品责任。一方面，产品责任适用严格的无过错归责原则；另一方面，市场主体明知算法系统存在缺陷与风险而未告知的，须承担惩罚性赔偿责任。这两方面都加重了市场主体未告知的法律后果，使算法决策告知义务具有刚性规范的性质。同时，现行消费者保护与产品责任相关规定可以提供现成的法律依据，立法成本同样很低。

〔63〕 参见刘东亮：《技术性正当程序：人工智能时代程序法和算法的双重变奏》，《比较法研究》2020年第5期，第72页以下。

〔64〕 相关新闻报道，参见《“单踏板”该为事故买单吗?》，《新能源汽车报》2022年12月5日第14版。

结 语

囿于价值定位的片面、技术发展的局限和规范性质的摇摆，算法透明理念及制度长期以来未得到法学界的一致认可。经过透明价值的法哲学思考、可解释技术的条分缕析与规范性质的类型化配置，笔者提出，算法透明制度不仅必要，而且可行，呈现出层次分明的价值内核与技术架构，可以被设计成由柔性规范、中性规范和刚性规范组成的多层次体系。

面对日新月异的技术发展与社会变革，单纯肯定或否定算法透明的观点会逐渐失去生命力，而本文尝试构建的多层次分析框架将历久弥新。与我国法学界近年流行的类型化或场景化分析进路类似，层次化分析框架同样强调分门别类，避免非黑即白，注重具体问题具体分析。与之不同的是，层次化分析更关注不同层次间的递进关系与耦合关系，探索多维度的层次划分与相互作用及其对规范配置的影响。在算法治理体系中，这种层次化的分析与治理进路还可以运用于解决算法准确性、算法隐私侵害、算法歧视等其他治理难题。上述难题同样暗藏着彼此制衡的治理目标，依赖难度各异的技术路线，需要不同性质的规范相互配合协调。多元的价值、多元的技术、多元的规范，构成层次化的多元框架，将为算法治理开辟一片新天地。

Abstract: The functional value, technical foundation and normative system of algorithmic transparency are worthy of discussion. At the value level, in addition to the two instrumental values, namely the value of improvement and the value of proof, algorithmic transparency also has such endogenous values as enhancing communication, rationality and trust in the algorithmic society and supervising the power of algorithms, which promote personal freedom and safeguard human dignity. At the technical level, explainable artificial intelligence has developed rapidly in recent years, forming technical levels such as model-centric transparency and subject-centered transparency, intrinsic transparency and post-event transparency, and observable, decomposable and simulated transparency, which provide a technical basis for opening the algorithmic black box. At the normative level, the current algorithmic transparency system swings left and right in the nature of norms, and the aforementioned value objectives, technical foundation and normative system should be coupled with each other to build a multi-level system composed of flexible norms, neutral norms and rigid norms.

Key Words: algorithm, the principle of transparency, explainable artificial intelligence, algorithmic black box