腾讯研究院 腾讯天衍实验室 TENCENT JARVIS 腾讯优图 Tencent AI Lab

可解释 AI 发展报告 2022

打开算法黑箱的理念与实践

可解释 AI 发展报告 2022

打开算法黑箱的理念与实践

腾讯研究院
×
腾讯天衍实验室
×
腾讯优图实验室
×
腾讯 AI Lab

目录

一 可解释 AI 概述

2016 年以来，以机器学习（machine learning），尤其是深度学习（deep learning）为代表的新一代人工智能技术不断朝着更加先进、复杂、自主的方向发展，这给经济和社会发展带来了新的变革性机遇。AI 应用迎来“物种大爆发”，日益渗透到各行各业和人类生活的方方面面，有望塑造新型的经济和社会形态。与此同时，科技伦理也日益成为了当前 AI 技术发展与产业应用中的“必选项”，各界纷纷探索 AI 伦理原则、框架、治理机制等。科技伦理的一个核心议题就是人工智能的透明度与可解释性（transparency and explainability）。2021 年 11 月，联合国 UNESCO 通过的首个全球性的 AI 伦理协议《人工智能伦理建议书》（Recommendation on the ethics of artificial intelligence），提出的十大 AI 原则就包括“透明性与可解释性”，即算法的工作方式和算法训练数据应具有透明度和可理解性。[1]

虽然并非所有的 AI 系统都是“黑盒”（black box）算法，并非比非 AI 技术、传统软件或人工程序更加不可解释，但就当前而言，机器学习模型尤其是深度学习模型往往是不透明的，难以为人类所理解的。未来，人工智能的持续进步有望带来自主感知、学习、决策、行动的自主系统。然而，这些系统的实际效用受限于机器是否能够充分地向人类用户解释其思想和行动。如果用户想要理解、信任、有效管理新一代的人工智能伙伴，人工智能系统的透明性与可解释性就是至关重要的。因此，近年来，可解释 AI（Explainable Artificial Intelligence，简称“XAI”）成为了 AI 研究的新兴领域，学术界与产业界等纷纷探索理解 AI 系统行为的方法和工具。

（一）机器学习模型的可解释性挑战

作为引领 AI 技术加速变革的重要法宝，机器学习是一把双刃剑。一方面，它可以帮助 AI 摆脱对人为干预和设计的依赖，凭借自身强大的数据挖掘、训练和分析能力，完成算法模型的自主学习和自我更迭，使得 AI 在学习思维上无限接近于人类大脑，也被认为是 AI 由弱人工智能迈向强人工智能形态的关键性因素。更深入地说，深度学习即深度神经网络算法（Deep neural network）是 21 世纪 AI 发展的“制胜法宝”。神经网络的特征在于，其无须经过特定的编程，

1 https://en.unesco.org/artificial-intelligence/ethics

就能自动从偌大的数据库中学习并构建自身的规则体系。这样的自动生成逻辑是算法工程师的福音，能够极大地解放他们的生产力，并且可以适用于更加多元化的应用场景，形成 AI 的自主学习、自我创造以及自动迭代机制。

另一方面，机器学习又日益暴露出 AI 在自动化决策（Automated decision-making）中的伦理问题和算法缺陷。实际上，如同一个硬币的两面，深度学习算法既有其先进、独特的更迭优势，又有着无可回避的难解释性和黑箱性。在深度学习领域，基于人工神经网络结构的复杂层级，在 AI 深度学习模型的输入数据和输出结果之间，存在着人们无法洞悉的“隐层”，深埋于这些结构底下的零碎数据和模型参数，蕴含着大量对人类而言都难以理解的代码和数值，这也使得 AI 的工作原理难以解释。因此，深度学习也被称为“黑盒”算法。这些所谓的“黑盒”模型可能过于复杂，即使是专家用户也无法完全理解。早在 1993 年，学者 Gerald Peterson 就指出，除非人类能够说服自己完全信任这项技术，否则神经网络算法将不会被应用于关键领域，而增进信任的核心在于使人类能够理解 AI 的内部运行原理。该论断直到今天都仍为人们所接受和认同。

因此，神经网络的可解释性问题被认为是近三十年来 AI 技术的重点攻关方向。同时，对神经网络算法的解释困境也使得机器学习陷入了 AI 的寒冬期。[2] 在此背景下，XAI 的命题被提出，旨在解决深度学习算法所带来的解释性问题。从可解释 AI 的缘起和发展进程来看，随着深度学习算法逐渐被人们奉为圭臬，XAI 成为深度学习的“开箱者”，逐渐走进公众的视野，也为人们带来了开启算法解释大门的“金钥匙”。

（二）可解释 AI 的意义

透明性与可解释性是对 AI 系统的基本要求，是实现其他伦理价值的必要前提，但也需要与隐私、安全等其他原则进行平衡。综合来看，透明、可解释 AI 具有以下意义：

第一，帮助用户增强对 AI 系统的信心与信任。对 AI 的自动决策提供一个解释，可以在很大程度上增进人们对 AI 系统的信心与信任，尤其是在预测的准确性与合理性方面。用户寻求解释的目的多种多样，包括深入学习、与社会良性互动、分配责任等。通过对算法决策的解释，一

2 Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3559477

方面能够保障公众基本的知情权和同意权，另一方面也有利于增强公众对 AI 产品的信心与信任。而且缺乏透明度可能阻碍 AI 系统的可用性。

第二，防止偏见，促进算法公平。可解释 AI 旨在对算法黑箱、算法失灵等问题作出回应，赋予用户对算法决策机制的知情权等，通过算法透明化机制倒逼开发者、提供者采取有效的措施防范算法歧视、决策偏差等问题，从而促进算法公平，实现科技向善。

第三，满足监管标准或政策要求。透明度或可解释性对于行使围绕 AI 系统的法律权利、证明产品或服务符合监管标准，以及帮助解决有关责任的问题具有重要意义。在满足监管标准的条件下，可解释 AI 将摆脱决策合法性的质疑，能够更加高效无误地执行特定事务，而无须担忧决策结果是否会引发政府的干预和制裁，从而提升决策的效率。此外，AI 可解释性对于确保问责（accountability）也至关重要，至少可以为质疑 AI 系统的输出、结果提供基础。

第四，理解和验证 AI 系统的输出，推动系统设计的改进。可解释性可以帮助开发人员探究、分析 AI 系统以某种方式产出某种结果的原因，并对 AI 系统的输出进行验证，进而帮助开发人员理解 AI 系统，做出正确的决定，在此基础上对 AI 系统做出改进。例如，在自动驾驶汽车中，了解系统为什么以及如何发生故障；在医疗保健领域，可解释性可以帮助医疗人员理解结果，研发人员可以跟踪系统故障等。总之，可解释性能够为研发人员提供更好的审计路径和问责机制，助力 AI 系统的升级更迭。

第五，帮助评估风险、鲁棒性和脆弱性。欧盟委员会设立的人工智能高级别专家组在 2019 年 4 月发布了《可信人工智能的伦理指南》，其中提出可信人工智能需要满足鲁棒性（robustness），即可信人工智能应当避免造成无意的损害或负面影响。因此，有必要对 AI 系统的运作程序进行风险评估，特别是当 AI 系统部署在一个全新的环境中，而 AI 的可解释性可以帮助开发人员了解系统是否容易以及可能会如何受到对抗性攻击等。

（三）可解释 AI 与透明度、问责制

对可解释 AI 的讨论往往离不开两个关联概念，也即透明度（transparency）和问责制（accountability）。概言之，透明度是可解释 AI 的主要目标，而可解释性是实现透明度的一种方式；问责制是可解释 AI 的必要保障机制，而可解释性主要是为了确保能够问责。联合

国《人工智能伦理建议书》也指出，透明性、可解释性与责任（responsibility）、问责措施（accountability measures）、AI 系统的可信性等密切相关。

具体而言，在 AI 中，透明度是指对外提供 AI 算法的内部工作原理，包括 AI 系统如何开发、训练和部署，以及披露 AI 相关活动，以便人们可以进行审查和监督。透明度并不一定意味着基础信息能被人们理解，其仅能保证信息的原始性和真实性。要想让人们理解系统的透明信息，还需要解释机制将信息予以描述和说明。总之，透明度的目的是为相关对象提供适当的信息，以便他们理解和增进信任。

需要特别说明的是，透明度并不代表要将所有系统信息或背后的数据集都予以公开，全部公开反而可能带来严重的安全风险，对于增进 AI 责任和打造 AI 信任机制几乎没有助益。例如，单纯披露源代码（source code）、训练数据集或单个用户的数据，对于人们理解 AI 系统决策机制并无助益，但却可能让 AI 系统陷入被滥用或被篡改的境地，也可能给用户隐私与企业的知识产权带来侵权风险。政策制定需要考虑透明度要求与其他重要目的（如效率、安全、隐私、网络安全等）的平衡，实现各种原则之间的友好协调。正因如此，欧盟、美国在立法上并没有要求科技公司披露 AI 系统的全部信息。联合国《人工智能伦理建议书》也指出，在存在对个人权利产生不利影响的严重威胁的情况下，透明度要求可能还包括共享代码或数据集。

关于可解释性与问责制，根据经合组织的说法，问责制是指让适当的组织或个人承担 AI 系统正常运作的责任的能力。问责制处于 AI 伦理层次结构的顶端，其他 AI 伦理目标都要为之作出贡献。在 AI 可解释的思路下，问责制要求 AI 具备一些理想的特征，如尊重人类价值和公平性、透明度、稳健性和安全性等原则。[3] 人类价值和公平性要求对 AI 系统进行灵活的、针对特定背景的解释。可以说，可解释 AI 的核心目标之一就是确保能够对 AI 系统进行问责。

总之，AI 的透明度、可解释性本身不是最终目的，而是实现其他目的（如问责）的手段和前提条件。因此，在设计可解释性与透明度要求时，监管者需要考虑其最终所要实现的目标是什么，在特定情境下 AI 可解释性要求如何更好地匹配这些目标，以及如何设计程序才能够更好地实现 AI 的可追溯和可追责，这些都是在追求可解释 AI 时需要予以特别考虑的重要问题。

3 OECD：Artificial Intelligence in Society，https://www.oecd.org/publications/artificial-intelligence-in-society-eedfee77-en.htm

（四）可解释 AI 的两个维度

对于 AI 的可解释性的定义，联合国《人工智能伦理建议书》指出，可解释性是指让人工智能系统的结果可以理解，并提供阐释说明。人工智能系统的可解释性也指各个算法模块的输入、输出和性能的可解释性及其如何促成系统结果。对于可解释性的具体认定，美国国家标准与技术研究院（NIST）提出了解释人工智能决策的四项基本原则，分别是：（1）人工智能系统应该为其所有输出结果提供相应的证据或理由；（2）系统应该对个人用户提供有意义 / 可理解的解释；（3）该解释正确地反映了系统生成输出结果的过程；（4）系统只在其设计的条件下运行，或者说系统对其输出结果应该有足够的信心，否则就不应该向用户提供决策。

一般而言，可以从两个角度对 AI 的可解释性进行理解，一是要理解算法本身，也就是要对 AI 的系统功能（System functionality）进行全局解释（Global Interpretability）；二是要理解算法的特定输出，也就是要对 AI 运作中的特定决策（Specific decision）进行局部解释（Local Interpretability）。从技术角度来看，全局可解释能够基于完整数据集上的依赖（响应）变量和独立（预测变量）特征之间的条件交互来解释和理解模型决策。[4] 局部可解释性更加关注模型中的单条样本或一组样本。通俗来讲，如果让 AI 对一张人脸识别图进行解释，前者能够回答“人脸是什么样的”，后者回答的是“为什么这是一张人脸”。因此，全局可解释立足于 AI 的整体功能进行解释，而局部可解释立足于 AI 的具体功能进行解释。从不同的维度进行解释，用户接受和理解到的信息也会有所差别。

1. 局部可解释

局部可解释性的关键在于对个案采用特定化、情景化的解释方法，让 AI 系统的用户群体能够充分接收并理解特定决策的生成路径。从要素组成来看，特定决策的局部可解释性应涵盖以下内容：决策依据的规则、主要参考因素、各类因素的权重占比、决策中参考或者引用的信息来源。[5] 在这些局部解释因素提供的支撑下，XAI 能够对特定决策下用户最为关心的问题进行解释和说明。例如，作出决策的主要影响因素是什么，改变模型中的某个因素是否会改变最终的决策，相似的模型案例得出不同决策的原因是什么，不同的模型案例得出相同的决策的原因是

4　A Survey of Methods for Explaining Black Box Models, https://arxiv.org/abs/1802.01933

5　Sandra Wachter, Brent Mittelstadt, Luciano Floridi: Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2903469

什么等。[6] 这些问题直接关系着 AI 系统的用户能否充分理解算法的决策机制，也影响着决策过程及结果的公平性。

2. 全局可解释

全局可解释性是为了解释模型的全局输出，需要训练模型了解算法和数据。从整体的维度来看，可解释 AI 在对具体个案做出特定决策时，也要基于 AI 本身的系统特征进行适用，所有特定化的决策都离不开 AI 的整体功用，包括 AI 的算法类型、自动决策程序、可控范围、相关数据库等基本要素。全局可解释性可以采取多种形式，从源代码和训练数据的交流，到算法功能的简单概述，例如谷歌对其搜索算法如何工作的解释。然而，在其他情况下，披露源代码可能对全局可解释性没有什么价值。[7] 全局可解释性可以通过对算法如何运作的描述来实现，经过解释后的 AI 可以回答用户对 AI 的各种问题，例如全局可解释的模型理论、学习路径以及训练数据的使用等，从源头上厘清 AI 的全局性难题，也包括针对某些特定的任务训练一个可解释的 AI 模型。

总之，从技术角度来看，可解释性的重点不仅在于 AI 系统是否可以被解释，或某一模型是否比另一模型更容易被解释，还在于 AI 系统是否能够提供特定任务或用户群体（例如一般用户或专家）所需的可解释性类型。一般用户、审查人员、监管者等不同主体有不同的需求。例如，当普通用户认为 AI 的决策存在错误或歧视时，可能会要求服务提供者就 AI 决策的理由和过程进行披露和解释，以便能够质疑、挑战 AI 决策。因此需要向用户提供明白易懂的、非技术语言的解释。对于专业人员而言，则需要更全面的，更多技术细节的解释，以便评估 AI 系统是否满足稳健、安全、准确等方面的一般性要求。

此外，英国信息专员办公室（ICO）与艾伦图灵研究所联合发布的指南《解释人工智能做出的决定》(Explaining decisions made with AI)，提出了 AI 的可解释性的六种主要类型，也具有一定的启发意义。该指南提出的可解释性的六个类型包括：[8]

6 Accountability of AI Under the Law: The Role of Explanation, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3064761
7 The Intuitive Appeal of Explainable Machines, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3126971
8 https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-ai/

（1）基本原理解释：

用一种可理解的、非技术化的方式，说明人工智能做出决定的原理。

（2）责任解释：

谁参与了人工智能系统的开发、管理和实施，以及与谁联系以对决定进行人工审查。

（3）数据解释：

在特定决策中使用了什么数据以及如何使用。

（4）公平解释：

在人工智能系统设计、部署的过程中，采取措施确保其决策大体是公正和公平的，不管个体之前是否被公平对待。

（5）安全性和性能解释：

在人工智能系统设计、执行的过程中，采取措施使决策和行为的准确性、可靠性、安全性和鲁棒性达到最大化。

（6）影响解释：

在人工智能系统设计、部署的过程中，考虑并且监控使用该系统可能带来的影响，包括对个人以及对社会的影响。

二 可解释 AI 发展趋势

回溯历史，自 2016 年起，世界各地的政府及各类非官方社会组织就开始极力呼吁加强 AI 的可解释性。美国作为全球 AI 战略的领跑者，从联邦到州政府、从政府到企业层面，制定了一系列的行业规范。如美国电气和电子工程师协会（IEEE）分别于 2016 年和 2017 年连续颁布了《人工智能设计的伦理准则》白皮书，重点强调了对人工智能和自动化系统应有解释能力的要求。另外，美国计算机协会、美国公共政策委员会在 2017 年发布的《算法透明性和可问责性声明》中提出了七项基本原则，其中包含着对 AI 可解释的要求，督促使用算法决策的系统和机构要主动对算法的过程和决策提供解释。此外，美国国防部积极支持关于寻求创造更多可解释的 AI 的研究计划。在英国，上议院 AI 委员会就认为，如果 AI 要成为社会中不可或缺且值得信赖的工具，那么对可解释 AI 的研究是非常有必要的。在欧盟，2019 年，欧盟出台《可信人工智能的伦理指南》，明确提出人工智能的发展方向应该是“可信 AI”（trustworthy AI），包含安全（security）、隐私（privacy）、透明（transparency）、可解释（explainability）等方面的要求。此外，欧盟呼吁应当进一步定义实现可解释性的途径。

（一）AI 的透明性和可解释性逐渐成为立法和监管关注的焦点

在欧盟，欧盟《数字服务法案》（尚未正式出台）升级了网络平台尤其是超大型网络平台（very large online platform）的注意义务，在算法透明度方面提出了新的要求。主要包括：（1）在服务条款中说明推荐算法系统的主要参数维度，以及用户修改或影响这些参数维度的选项；（2）在软件界面上提供易用的功能选项，方便用户选择或修改内容的呈现方式，包括至少向用户提供不基于其个人画像的选项，即允许用户退出（opt-out）个性化推荐。欧盟《人工智能法案》（尚未正式出台）也针对高风险 AI 系统提出了多项合规要求，其中就包括透明度及向用户提供信息，即确保高风险 AI 系统的运作足够透明，向用户提供使用说明（instructions of use）等。此外，有限风险 AI 系统需要遵守一定的透明度规则，包括向用户披露 AI 应用的使用情况等。此外，英国的《在线安全法案草案》（Draft Online Safety Bill）要求平台服务提供者就其所提供的每一项服务准备年度的透明度报告。加拿大的一项法案也要求在线科技平台建立标准化的透明度报告机制。美国曾在 2019 年推出了算法问责法案，但未能通过；现在准备

推出更新版的算法问责法案，拟要求在医疗、教育、住房、招聘等领域中使用自动化决策系统需要进行透明度报告（reporting）。

在我国，相关立法开始对人工智能算法应用的透明度和可解释性提出要求。例如，《个人信息保护法》第 24 条要求确保自动化决策的透明度和结果公平、公正，以及在自动化决策对个人权益有重大影响时个人可以要求提供说明。网信办等九部委出台的《关于加强互联网信息服务算法综合治理的指导意见》，将“算法应用公平公正、透明可释”作为算法治理的基本原则之一。在具体要求上，《指导意见》要求推动算法公开透明，督促企业及时、合理、有效地公开算法基本原理、优化目标、决策标准等信息，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。《互联网信息服务算法推荐管理规定（征求意见稿）》也要求算法推荐服务提供者优化算法推荐相关规定的透明度与可解释性，向用户公示算法推荐服务的基本原理、目的意图、运行机制等。

（二）对政府公共部门使用的 AI 系统提出较高的透明度与可解释性要求

随着 AI 技术的广泛普及和深度应用，AI 正在重塑政府部门和私人领域的决策方式，已经并将继续对人们的生活产生巨大影响。尤其在政府部门，法规政策、行政命令等政府行为对人们权益的侵入性更强，AI 在政府领域的使用也需要更高的应用限制和解释标准。因此，在政府使用 AI 系统的情况下，对可解释性要求普遍高于在商业领域中使用 AI 系统。现阶段，美国、荷兰等政府明确使用 AI 系统必须实现可解释性要求。在法国，相关立法对政府部门使用的 AI 系统的可解释性要求作出了详细说明，包括：算法处理对相关决定的贡献程度和方式；用于处理的数据及其来源；在个别处理过程中使用的参数及其权重；以及由处理产生的操作。法国相关立法要求全局可解释性以及局部可解释性，而且必须在一开始就告知用户该处理涉及算法，否则行政机关的决定无效。[9]Ada Lovelace 研究所等智库呼吁对政府公共部门使用的算法决策系统进行透明度登记（transparency register），提供算法系统相关的信息，以便公众、媒体、学术、社会组织等可以进行监督。[10] 加拿大则针对政府公共部门使用的算法提出了算法影响评估（algorithmic impact assessment）机制。

9 French Code of Relations between the Public and the Administration

10 https://venturebeat.com/2021/12/08/the-u-k-s-new-ai-transparency-standard-is-a-step-closer-to-accountable-ai/

此外，英国中央数字数据办公室（CDDO）联合数据伦理与创新中心（CDEI）针对政府公共机构使用 AI 算法系统推出了统一的算法透明度标准，使英国成为世界首个发布算法透明度跨政府标准的国家。[11] 新发布的算法透明度标准包括两个层次：第一层包括对算法工具的简要描述，包括如何使用以及为何使用等；第二层包括算法工具如何运作的更详细信息，以及用来训练模型的数据集、人类监督的程度等。同时，英国政府也承诺，该标准将率先由多个公共部门进行试点，并在反馈的基础上进一步发展。开放政府合作伙伴（Open Government Partnership）联合阿达 · 洛芙莱斯研究所（Ada Lovelace Institute）和 AI Now 研究所开展了一项关于公共部门的算法问责制的研究，其也特别强调了公共部门的强制性报告义务在提高算法透明度方面的作用。[12] 阿达 · 洛芙莱斯研究所（Ada Lovelace Institute）副主任伊莫金 · 帕克（Imogen Parker）也表示：“公共部门算法工具使用的公开透明，对数字政府可信化来说意义重大。我们期待看到试验、测试和迭代，期待政府部门和公共部门机构发布完整的标准。”此外，英国关于未来数据保护制度的公众问询，也提议针对政府公共部门使用的自动化决策算法建立强制性的透明度报告义务。这些无不表明提升政府公共部门使用算法的透明度的清晰方向。

（三）对商业领域的 AI 系统在可解释性方面避免作“一刀切”要求

AI 系统在商业领域具有更加广阔的应用前景，小到手机的人脸或语音识别，大到汽车的自动驾驶技术，AI 与传统商业的结合催生出各类智能化的产品和服务。商业领域的 AI 系统在业务范围、功能类型、用户群体上都有所差异，这意味着相关的可解释性标准需要考虑行业与应用场景的差异，不宜提出“一刀切”要求，从而在提升 AI 商业价值与保护公众利益之间达成平衡。

其一，考虑 AI 可解释的不同类型，且考虑不同应用场景的差异性。对商业领域中 AI 系统的可解释性要求主要有两种类型：一类是要求对结果进行原因解释，另一类是要求对算法模型逻辑进行解释。美国法律对 AI 可解释性要求就是要求对结果进行原因解释，主要体现在《公平信用报告法》（FCRA）[13] 和《平等信用机会法》（ECOA）[14]，这两部法案都规定了“不利行动告知”（adverse action notice）条款。这些条款要求有关对于消费者不利的行动告知，无论基于信用评估的决策是否基于 AI 系统的算法模型，都必须包括原因说明，如拒绝信贷或其他基于信用的

11 https://www.gov.uk/government/collections/algorithmic-transparency-standard
12 https://www.opengovpartnership.org/wp-content/uploads/2021/08/algorithmic-accountability-public-sector.pdf
13 15 U.S.C. § § 1681–1681x（2012）
14 15 U.S.C. § § 1691–1691f（2012）

结果的原因说明。可见，基于结果的解释就是描述与决定相关的事实，但不描述决策规则本身。

而欧盟地区的立法则要求对算法模型逻辑进行解释。欧盟《通用数据保护条例》（GDPR）规定，在任何对其有法律影响或类似重大影响的自动决策中，数据主体可以要求获得“关于所涉及的逻辑的有意义的信息”，但 GDPR 并没有要求对特定自动化决策的结果进行解释。更进一步而言，GDPR 要求对算法模型进行功能描述，对管理决策的规则进行足够的描述，以便数据主体能够维护其在 GDPR 和人权法下的实质性权利；基于逻辑进行解释，描述一项决定背后的推理，而不仅仅是决定的相关输入。欧盟于 2021 年 4 月发布的《人工智能法案》则对高风险 AI 系统提出了较高的透明度要求和信息披露义务，第十三条要求系统开发者应以合适的数字格式或者其他说明向用户提供相关联（relevant）、可访问（accessible）、可理解（comprehensive）的信息，并对需要披露的信息作了完整的说明，包括系统开发者的基本信息、高风险系统的性能特征、监督措施以及维护措施等等。[15]

此外，《欧盟平台商户关系条例》（EU Platform to Business Regulation）规定了在线平台和搜索引擎对排名算法的解释义务，在线平台或搜索引擎必须对影响平台排名的“主要参数”进行“合理描述”，包括“一般标准、流程、纳入算法的具体信号或其他与排名相关的调整或降级机制”。在线平台服务和搜索引擎不需要披露其排名机制，包括算法的详细运作情况，也不需要披露商业秘密，但描述必须披露基于所使用的排名参数的相关性的实际数据；解释必须以“通俗易懂”的语言说明，并允许企业用户在使用服务的背景下充分理解排名的功能；在线平台和搜索引擎给出的“合理描述”须是有用的，即它应该帮助企业用户改进其商品和服务的展示；解释的内容应几乎完全与用户的可理解性和实用性相联系。

其二，根据商业 AI 的风险等级与影响大小，针对不同风险等级的 AI 系统建立不同标准，避免作一刀切式的监管要求。如前所述，GDPR 虽然也涉及算法解释相关要求，但 GDPR 规定的算法解释仅仅针对产生法律效果或类似的重大效果（如信贷、就业机会、健康服务、教育机会等）的、不存在人类干预的完全自动化决策，体现了分级分类分场景的监管思路。此外，欧盟在 2020 年发布的《人工智能白皮书》中提出，并非对所有的 AI 应用都实施监管，只针对“高风险”的 AI 应用进行监管。欧盟《人工智能法案》在《人工智能白皮书》的分级分类理念基础上，进一步针对高风险 AI 系统提出了监管要求。“高风险类 AI”是指可能对人类生活和

15 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

权利产生重大影响的 AI 系统，其应用包括以下领域：关键基础设施（如交通），教育或职业培训，产品的安全组件，就业、员工管理和自雇佣机会，基本的私人和公共服务（如信用评分），可能会干扰人们基本权利的执法活动，移民、庇护和边境管制，司法和民主程序。这些场景涉及人类的基本社会经济权利，相应监管要求也相对更高。《人工智能法案》针对高风险 AI 系统提出的监管要求包括技术记录、记录保存、透明度及向用户提供信息（运作足够透明，提供使用说明），以及人类监督等，但并未要求 AI 算法决策结果层面的可解释性。此外，对于有限风险 AI 系统提出了一定的透明度规则：（1）与人类交互的 AI 系统如聊天机器人需要向用户披露其身份；（2）情形识别系统或生物特征分类系统需要向用户披露；（3）Deepfake 应用需要披露、标注内容的合成属性。最后，避免监管最小风险（minimal risk）AI 应用，绝大部分 AI 系统属于这个范畴，无须承担新的法律义务，遵守既有监管规定即可。国外立法对 AI 系统在可解释性等方面的监管要求，一方面区分公共领域的 AI 应用与商业领域的 AI 应用；另一方面从具体行业与应用场景出发，采取基于风险的监管思路，而非进行一刀切立法，体现了平衡创新与伦理的治理理念。

（四）行业积极探索可解释 AI 的技术解决方案

随着 AI 技术与各个行业的结合趋于紧密，作为实现可信 AI 的重要组成部分，AI 可解释性成为行业发展 AI 技术的重要方向。科技企业纷纷加大研发投入，积极探索、落地可解释 AI 的行业方案。例如，微软、谷歌、IBM 等硅谷科技公司，通过打造各具特色的可解释 AI 相关工具与服务，致力于提高 AI 算法的可解释性和决策模型的透明性。

受益于相关学术研究、监管要求、公众期待与行业探索，围绕可解释 AI 的机制、工具、服务等不断涌现，推动可解释 AI 不断向前发展，具有代表性的为谷歌公司的模型卡片机制（Model Cards）、IBM 的 AI 事实清单（AI Fact Sheets）机制以及微软的数据集数据清单（Datasheets for Datasets）等。此外，谷歌、微软、IBM 等科技公司为了更好地帮助开发者打造可解释的 AI 模型，还打造了各具特色的可解释工具与服务，如谷歌的 Explainable AI 工具包，帮助开发者实现可解释 AI 目标。未来，随着越来越多的科技公司与创业公司布局可解释 AI 等 AI 伦理研究与应用落地，将能极大助力可信 AI 与 AI 信任目的之实现。

三 可解释 AI 的行业实践

随着人们对可解释问题的重视程度的不断上升，对该问题解决方案的渴求也越来越强烈。一方面，这使得相关研究的数量激增，理论的探讨层出不穷；而另一方面，则体现在越来越多的、可实践的可解释性工具被发布出来。早在 2016 年起就开始出现一些可解释性的工具，来解释机器学习分类模型为何得到如此的结果（ELI5）。而后，越来越多的可解释性工具开始囊括了更多的功能，可以同时对不同的统计机器学习模型和深度学习模型进行解释，包括一般的泛线性模型、集成学习模型、图像识别模型以及自然语言处理模型等。而近年来头部的人工智能公司，包括微软、谷歌等，更是推出了更加强大与丰富的可解释性工具，囊括了诸多可诠释（Interpretable）方法与可解释（Explainable）方法，为实际面临的可解释性问题的解决提供了巨大的帮助。

行业层面来看，可解释 AI 并非是算法模型中的某个独立环节或者是某类具体工具，构建一个可解释 AI，往往需要从算法模型生命周期中的各个步骤进行介入。一般而言，算法的完整生命周期大概包括以下五个环节：

在诸多环节中，不同的环节有不同的可解释方法。比如，微软的 Datasheets for Datasets（数据集数据清单）就主要应用于数据准备阶段，关注机器学习模型训练所使用的数据集是否存在偏见的可能性；而谷歌的 Model Cards（模型卡片）则聚焦于部署和监控环节，就算法模型本身的性能表现、局限性等指标提供解释。

实际上，业界更为主流的框架为“事前可解释性”（Ante-hoc）以及“事后可解释性”（Post-hoc）的分类。前者所涵盖的算法模型又被称为“内在可解释模型”，一般结构足够简单，可以通过观察模型本身来理解模型的预测过程。事后可解释方法，是给定训练好的模型和一个或一

批数据，尝试理解为什么算法模型要进行某些预测。具体的解释方法又可分为全局可解释方法和局部可解释方法。目前业界流行的大部分 AI 可解释机制与工具都属于事后可解释的范畴。

就目前而言，谷歌、IBM、微软三家科技公司在可解释 AI 行业的实践走在前列，通过不断创新探索出了各具特色的 AI 可解释性机制、工具与服务等，其中具有代表性的 XAI 产品分别为谷歌的模型卡片、IBM 的 AI 事实清单以及微软的数据集数据清单。三者的宗旨都是为了实现 AI 算法的可解释化，但在原理、功能及用途上有所区别。

（一）谷歌模型卡片（Model Cards）机制

围绕 XAI 的主线，谷歌推出了一系列技术举措。模型卡片是谷歌公司于 2019 年推出的一项具有算法解释功能的技术，也是谷歌在可解释 AI 领域的最新实践成果之一，并于 2021 年升级、开源了 Model Card Toolkit，以帮助机器学习从业者更容易地推进模型透明度报告。[16]

1. 模型卡片的运作原理

Model Cards 是一种情景假设分析工具，作用在于能够为 AI 的算法运作提供一份可视化的解释文档。该文档能够为使用者阅读，使其充分了解算法模型的运作原理和性能局限。从技术原理上看，模型卡片设置的初衷是以通俗、简明、易懂的方式让人类看懂并理解算法的运作过程，其实现了两个维度的“可视化”：一是现实算法的基本性能机制；二是显示算法的关键限制要素。

为了更好地介绍模型卡片的功能，在此以日常生活情景为例作以说明。正如我们在食用食物之前会阅读营养物质成分表，在路上行驶时会参考各种标志牌来了解道路状况，Model Cards 所扮演的角色，便是算法的“成分表”与“标志牌”。这反过来也提醒我们，即便对待食物或驾驶都如此谨慎，算法在我们的工作与生活中扮演着愈发关键的角色，我们却在没有完全了解它的功能与原理的情况下就听从其安排。算法在什么条件下表现最佳？算法有盲点存在吗？如果有，哪些因素影响了它的运作？大部分情况下，我们对这些问题都一无所知。在某种程度上，人之所以无法与算法“交流”，是因为后者的复杂原理，更进一步说，这是由于人与算法或更广义的 AI 采用不同的“语言”。人类使用高阶语言进行思考和交流，比如我们在形容一个事物

16 https://ai.googleblog.com/2020/07/introducing-model-card-toolkit-for.html

时往往会用颜色、大小、形状等维度的形容词。而算法关注低阶要素，在它的“视阈”里，一切元素都被扁平化为数据点，方便其考察不同特征属性（Feature Attribution）的权重。基于此，模型卡片能够辅助 AI 用户更好地审视和理解算法的运行路径，确保用户对算法决策的知情和理解。

图 1
图像识别算法运作原

以图像识别为例，对于算法来说，一幅图像中的每个像素都是输入要素，它会关注图片中每一个像素的显著程度并赋予相关数值，以此作为识别的依据。对于人来说，就显然就不可能用“第五个坐标点的数值是 6”这样的方式来进行判定。而 Model Cards 就是以人类能够看懂的方式来呈现算法的运作原理，它实现了两个维度的“可视化”：显示算法的基本性能机制；显示算法的关键限制要素。换言之，模型卡片主要回答了这样一些问题：目标算法的基本功能是什么？该算法在什么情况下表现最好？何种因素阻碍着算法的运作？这些内容的可视化帮助使用者有效利用算法的功能，并避免其局限性。从技术角度来看，模型卡片通过对算法的各项运行数值进行分析和判定，揭示算法的真实性能和局限因素，从而达到解释的效果。

2. 模型卡片的应用场景

这项诞生于 2019 年底的技术尚未得到大规模落地应用。但谷歌在其主页上提供了关于模型卡片应用的两个实例“人脸识别（面部检测）”和“对象检测”，以展示它的运作原理。在人脸识别为例，模型卡片首先提供的是“模型描述”（Model Description），即算法的基本功能。根据示例，可以看到人脸识别算法的基本功能就是“输入”（照片或视频）、“输出”（检测到的每个面部及相关信息，如边界框坐标、面部标志、面部方向以及置信度得分等）。

Input: Photo(s) or video(s)

Output: For each face detected in a photo or video, the model outputs:

- Bounding box coordinates
- Facial landmarks (up to 34 per face)
- Facial orientation (roll, pan, and tilt angles)
- Detection and landmarking confidence scores.

No identity or demographic information is detected.

Model architecture: MobileNet CNN fine-tuned for face detection with a single shot multibox detector.

图 2
人脸识别模型卡片

首先，“性能”部分显示了识别算法在各种变量下的表现，例如面部大小和面部朝向，以及人口统计学变量（如感知肤色、性别和年龄等）。模型卡片从与算法训练数据不同的数据源中提取评估数据集，以此有效检测算法的真实性能。

其次，“局限性”列举了可能影响模型性能的因素，比如脸型大小（距离相机较远或瞳孔距离小于 10px 的面孔可能无法被检测）、面部方向（眼、鼻、口等关键的面部标志应处于正面）、灯光（照明不良的脸部可能无法检测）、遮挡、模糊、运动等，这些因素会影响人脸识别的效果。

模型卡片通过提供“算法成分表”的方式，向研究者或使用者展示算法的基础运行原理、面对不同变量的性能和局限所在。其实，对于模型卡的想象力远可以超越谷歌提供的两个案例，其他算法模型也可以采用模型卡片对性能进行分析及展示，比如用于语言翻译的模型卡片可以提供关于行话和方言的识别差异，或者测量算法对拼写差异的识别度。

谷歌表示，模型卡片的目的是帮助开发人员就使用哪种模型以及如何负责任地部署它们做出更明智的决定。目前，模型卡片的主要应用场景是谷歌云平台上的 Google Cloud Vision，后

者是谷歌推出的一款功能强大的图像识别工具，主要功能就是学习并识别图片上的内容。Google 利用在大型图像数据集上训练的机器学习模型，开发人员可以通过调取这个 API 来进行图片分类、以及分析图像内容，包括检测对象、人脸以及识别文字等等。而模型卡片则为 Google Cloud Vision 面部检测和对象检测功能提供了解释文档。这些功能对于各个行业的从业人员而言都有所裨益，能够在特定使用场景下赋予人们更加全面的理解涵义。

例如，对于技术人员来说，可以借助模型卡片来进一步了解算法的性能和局限，从而能够提供更好的学习数据，改善方法和模型，提高系统能力。

再如，对于行业分析师和媒体记者来说，可以根据模型卡片了解算法，从而更容易向普通受众解释复杂技术的原理和影响。而随着与模型卡片类似的技术思路得到更广泛开发和应用之后，更可以进一步使普通人从算法的透明性中获益。

模型卡片甚至可以帮助发现并减少算法偏见、算法歧视等问题。在基于人脸识别的犯罪预测系统中，算法在不同人群的识别上是否表现一致，还是会随着肤色或区域特征的改变而产生不同的结果？模型卡片可以清晰地展现这些差异，让人们清楚算法的性能及局限所在，并且鼓励技术人员在开发过程要重点考虑这些影响因素。

总体而言，以模型卡片为代表的“可解释性 AI”更像是一种对话方式。它不仅促成技术与技术人员之间的对话，而且也促成了专业人士与普通人的对话。

（二）IBM 的 AI 事实清单（AI Fact Sheets）机制

由 IBM 研究院研发的 AI 事实清单（AI Fact Sheets）是一项呈现算法模型重要特性的自动化文档功能。AI 事实清单的开发者认为，算法模型信息的标准化和公开化，有助于增强不同类型的使用者对模型的理解和信任，并且能够以此避免因算法模型不透明而导致的系列问题。正如 IBM 研究院的研究人员 Ritu Jyoti 所言：“通过对人工智能进行适当的管理，我们可以防止不良结果的发生，比如接受不适当或未经审核的数据的训练，或者是在性能上出现意外变化，而导致模型在无意中存在的偏见。”[17]

17 https://searchenterpriseai.techtarget.com/news/252493368/New-updated-IBM-AI-tech-bring-better-NLP-NLU-to-business-users

1.AI 事实清单的主要功能

基于上述目的，AI 事实清单在设计思路上同样类似于食品的营养标签或者家用电器的参数表，其功能是提供有关人工智能模型基本特征的重要信息，比如模型的目的、预期用途、模型性能、数据集等等。根据 IBM 官方的介绍，在 IBM Cloud Pak for Data 的基础上，AI 事实清单进一步完善并添加了新的功能，从而拓展了应用范围和可靠性，包括更直观的用户界面、增强治理和安全性以及联合学习功能，在确保数据隐私和安全的前提下，支持基于分布式数据集的模型训练。目前，AI 事实清单能够实现以下主要功能：

其一，策略创建（Policy Creation）： AI 事实清单的策略创建功能，允许使用者自定义在 AI 模型上所要采集及跟踪的信息，覆盖模型测试、训练、部署和评估等环节。它还可以判定哪些数据可以使用，哪些数据不能使用，以及哪些规定及公司政策需要考虑，也可以判定谁可以使用模型，用于什么目的以及应该如何运行等。

其二，自动数据采集（Automated Data Capture）： 一般而言，记录 AI 模型的性能需要大量的时间和资源，并且容易导致时效性差，或者生成的报告根本与模型无关等问题。而 AI 事实清单能够帮助使用者连续、自动地采集整个 AI 生命周期中（通过“策略创建”功能设定）的模型事实，包括模型表现（model performance）、准确性（accuracy）、公平性（fairness）、鲁棒性（robustness）以及可解释性（explainability）。

其三，自动报告（Automated Reporting）： 作为一种自动化文档功能，通过自动数据采集，AI 事实清单能够提供关于模型性能以及其他自定义指标的实时报告，并且基于不同用户的需求量身定制、输出可解释报告。

基于不同用户的需求和喜好进行个性化定制，也是 AI 事实清单最大的特性之一。它收集到的信息可以因行业和用例而有所差异，最终提供的报表或者视图也可以自定义。AI 事实清单机制可以提供给不同领域的研究者、企业主、数据科学家基于其特定需求的独特报告，从而能够实现跨不同级别、跨不同领域的技术专业知识进行协作，并且能够满足不同的透明度与合规要求。这也为 AI 可解释工作带来了启发，即可解释性并非铁板一块，而是取决于用户的具体需求。不同的用户群体需要不同类型的信息。比如数据科学家和普通的算法产品用户需要的信息显然是不一样的，而不同的 AI 应用程序或用例也意味着不同的信息需求。

2.AI 事实清单的基本原理

在关于 AI 事实清单的原始论文中，研究者提供了两个示例的 AI 事实清单，类别包含模型目的陈述、基本性能、安全性等，涵盖了服务开发、测试、部署和维护的各个方面：从有关服务训练数据的信息，到基础算法、测试设置、测试结果和性能基准以及服务的维护和再训练方式等等。[18]

论文中展示的 AI 事实清单的项目示例包括：

- 服务输出的预期用途是什么？
- 该服务运用了哪些算法或技术？
- 服务测试了哪些数据集？(提供指向用于测试的数据集的链接，以及相应的数据表。
- 描述测试方法。
- 描述测试结果。
- 是否知道使用该服务可能导致的偏见，道德问题或其他安全风险的例子？
- 服务输出是否可以解释和 / 或解释？
- 对于服务使用的每个数据集：是否检查了数据集是否存在偏倚？采取了哪些措施来确保其公平性和代表性。
- 服务是否实施并执行任何偏见和补救措施？
- 对看不见的数据或具有不同分布的数据的预期性能是什么？
- 是否检查了该服务是否具有对抗攻击的鲁棒性？
- 上次更新模型的时间是什么？

18 https://arxiv.org/abs/2006.13796

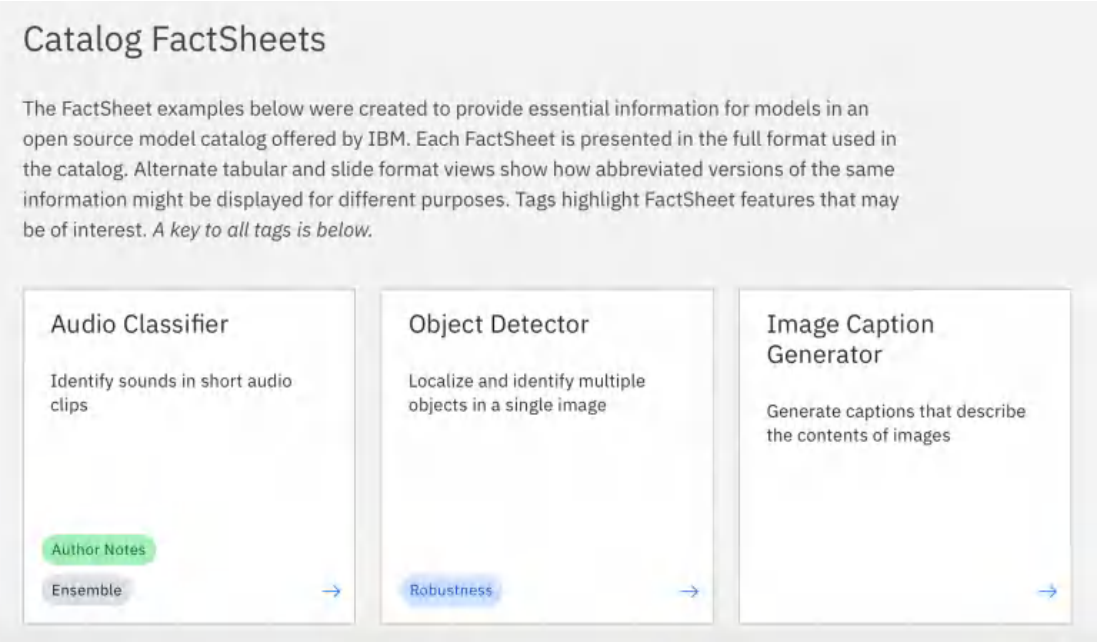

图 3 IBM AI 事实清单示例

以 AI 事实清单为代表的自动化文档是增强 AI 可解释性的重要方式，它能够以一种清晰明了的方式，作为技术人员与使用者的沟通介质，从而能避免许多情形下的道德和法律问题。AI 事实清单并不试图解释每个技术细节或公开有关算法的专有信息，它最根本的目标是在使用、开发和部署 AI 系统时，加强人类决策，同时也加快开发人员对 AI 伦理的认可与接纳，并鼓励他们更广泛地采用透明性可解释文化。

在过去，为了部署并扩展可解释 AI，开发者必须在 AI 的整个生命周期内，通过建立可信的模型，记录部署的 AI 模型，以便向监管机构和其他相关方进行报告。但整个过程需要耗费大量的时间和资源，而且这些文档通常是以临时、手动和不完整的方式构建的，因此效率并不高。

而 AI 事实清单通过自定义的策略创建，自动、准确地从 AI 模型采集信息，并且自动生成报告文档，这将有助于使用者为合规做好准备，降低管理风险。除了实用角度，AI 事实清单也将有助于提高透明度，增强对人工智能的信任。

回到 AI 事实清单的原始论文中，我们更能够看清研究者设计这一工具的基本思路和设想。在最初的设计中，AI 事实清单是以“供应商合格声明”（SDoc，Supplier's Declaration of Conformity）为基础原则的。SDoC 是表明产品或服务符合标准或相关技术法规的文件，供应

商需要提供符合制定要求的书面保证。一般而言，SDoC 通常是自愿的，并且报告的相关测试是由供应商本身而不是第三方进行的。这将自我声明与强制性认证（必须由第三方进行的测试）区分开来。行业参与而非政府监管，是 AI 事实清单最初的设想。

另外，在研究者的最初设想中，AI 事实清单需要由供应商提供描述性的大量细节，完成会相对费力，因此研究者希望以安全可审核的方式，将大多数基础性信息作为 AI 服务创建过程中的一部分进行填充。事实清单将首先被创建并与服务相关联，但是可以在不删除先前信息的情况下进行扩充，即从更多测试中添加结果，但无法删除结果。这意味着，对服务所做的任何更改都将为新模型创建一个新版本的事实清单。通过将已部署的指标与开发过程中看到的指标进行比较，并且在检测到意外情况时采取适当的措施，可以根据此信息来更准确地监视已部署的服务。

研究者设想的 AI 事实清单的生成流程是：供应商将自愿填充并发布事实清单，其中涵盖模型的关键信息，作为一种竞争优势，以保持自身在市场上的竞争力。随着 AI 服务市场的发展，最终可能会出现由第三方测试和验证的，由实验室、相关服务和工具组成的生态系统。“尽管制作事实清单最初将是 AI 服务生产者的额外负担，但我们希望 AI 服务消费者的市场反馈会鼓励这种创造。由于同行的压力要遵守，因此事实清单可能成为事实上的要求，类似于电器能源效率的能源之星标签。”

“它们将有助于减少供应商和消费者之间的信息不对称，因为消费者不了解服务的重要属性，例如服务的预期用途，性能指标以及有关公平性，可解释性和安全性信息。特别是，许多企业中的消费者缺乏评估市场上各种 AI 服务所需的专业知识，不明智的选择或不正确的选择可能会导致业务绩效欠佳。而通过创建 AI 事实清单，供应商可以通过赢得消费者的信任来获得竞争优势。”[19]

研究者也提到，尽管 AI 事实清单能够使 AI 服务生产者提供相关信息，从而使受过教育的消费者做出明智的决定，但消费者仍可能无意或恶意地将服务应用于预定目的以外的目的。AI 事实清单无法完全防止此类使用，但可以构成服务级别协议的基础。

19 https://arxiv.org/abs/2006.13796

在 AI 可解释性这一条道路上，除了 AI 事实清单，IBM 显然走得更远。2020 年 7 月，IBM 发布了名为“加速实现 AI 透明化的道路”的愿景，其中提供了有关公司和决策者如何推进 AI 可解释性发展和部署的详细信息与建议。此前，IBM Research 也推出一系列可解释 AI 的开源工具包，包括 AI Fairness 360，AI Explainability 360 和 Adversarial Robustness Toolbox 等等，更进一步地助力可解释 AI 的行业发展。

（三）微软的数据集数据清单（datasheets for datasets）机制

业界当前关于可解释 AI 的研究进展，集中在数据集和模型构建两个主要环节。前文中提及的 AI 事实清单与模型卡片，是综合性的可解释工具，遍布在 AI 生命周期的各个环节。而 Datasheets for Datasets（数据集数据清单）则是主要聚焦于算法训练数据集的可解释工具。

数据在机器学习和深度学习中扮演着至关重要的角色，几乎所有的机器学习模型和深度学习模型都使用数据集进行训练，其中不乏有诸如 ImageNet 或 COCO 这样的公共数据集，以及其他私有数据集。这些数据集的特征将从根本上影响模型的行为。如果数据集本身有部署语境，那么在脱离语境的场景下使用就不太可能表现良好，甚至会产生某些不可预测性的问题；如果数据集本身就存在一些“偏见”，那么势必也会造成一些不良后果。比如在 2020 年，超分辨率算法 PULSE[20] 将美国前总统奥巴马的肖像图转化为白人男性的形象，这起乌龙事件的主要“凶手”，就是该神经网络的训练所使用的 Flickr-Faces-HQ（FFHQ，人脸图像数据集），而这个数据集中大部分的图像素材都是白人照片。

当算法模型被应用于刑事诉讼等风险领域，以及关键基础设施、财政系统或者人员招聘时，这种不匹配可能导致严重的后果。问题就在于，尽管数据集对算法很重要，但是相关领域几乎没有用于记录机器学习数据集的标准化流程。世界经济论坛曾经发出倡议，所有主体都应该记录出处，创建和使用机器学习数据集以避免歧视的发生。这就是可解释工具 Datasheets for Datasets（数据集数据清单）推出的背景。在电子行业中，无论一个电子组件有多简单或复杂，都会附带有描述其工作特性、工作标准、测试结果、推荐用途以及其他信息的数据表。通过这些数据表，使用者在购买前就能够充分了解零件的相关信息，从而避免零件滥用，出现问题后也能够厘清责任归属。

20 https://thegradient.pub/pulse-lessons/

Datasheets for Datasets（数据集数据清单）就是算法训练数据集的“数据表”，也有媒体将它称为是数据集的“食物营养标签”。[21] 其真实用意在于，为每个数据集都随附一个“数据表”，这个数据表将记录其动机、组成、收集过程、推荐用途等等。通过数据表，使用者比如算法开发人员就能够了解他们所使用的数据的优势和局限性，并且防范偏见和过度拟合等问题。由于数据表的存在，也加强了数据集的生产者和消费者对数据源进行思考的可能性，它能够使人意识到，“数据”并不是真理来源，而是一种需要仔细审视和维护的资源。

Datasheets for Datasets 的原始论文，对这一可解释工具的原理及应用进行了深度探讨。[22] 基于研究者的原有想法，Datasheets for Datasets 旨在解决两个层面的关键需求，分别来自数据集创建者和数据集使用者。对于数据集创建者而言，Datasheets for Datasets 的主要目的是鼓励对数据集的创建、分发和维护过程进行思考 / 检索，包括其中可能涉及的任何潜在假设、潜在风险或危害，并着力提高数据集的透明度。而对于数据集的使用者而言，Datasheets for Datasets 的主要目标是确保他们获得所需的信息，并且制定有关数据集使用的恰当决策。

某种程度上，这是一体两面的一组关系。数据集创建者提高数据集的透明度，是使用者充分了解情况的必要条件；在充分了解的情况下，使用者才能基于自身任务需求选择合适的数据集，并且避免无意间的滥用。

除了这两个主要的利益相关者之外，Datasheets for Datasets 对于政策制定者、消费者权益的倡导者以及数据被包含在数据集中的个人等多方主体而言，也都意义深远。除此之外，Datasheets for Datasets 也能够达成一些次要目标：比如提高机器学习结果的可重复性；无须访问数据集，研究者和从业人员就可以使用数据表中的信息来重新调整、构造数据集。

根据论文的解释，Datasheets for Datasets 要求数据集或 API 的提供者须提供一份数据表，以解决一系列标准化问题（每个标准化问题框架下面，又有十几个相关的小问题，此处仅展示问题框架。完整问题列表可参见原始论文）：

21 https://venturebeat.com/2018/05/02/datasheets-could-be-the-solution-to-biased-ai/
22 https://www.microsoft.com/en-us/research/uploads/prod/2019/01/1803.09010.pdf

创建数据集的动机

Motivation for Dataset Creation

数据集的组成

Dataset Composition

数据收集过程

Data Collection Process

数据的预处理过程

Data Preprocessing

数据分发情况

Dataset Distribution

数据维护情况

Dataset Maintenance

法律和伦理考虑

Legal & Ethical Considerations

Prototypes of Datasheets for Datasets

A Database for Studying Face Recognition in Unconstrained Environments — Labeled Faces in the Wild

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should not be used?

The LFW dataset can be used for the face identification problem. Some researchers have developed protocols to use the images in the LFW dataset for face identification.[2]

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

Papers using this dataset and the specified evaluation protocol are listed in http://vis-www.cs.umass.edu/lfw/results.html

Who funded the creation of the dataset? If there is an associated grant, provide the grant number.

The building of the LFW database was supported by a United States National Science Foundation CAREER Award.

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Each instance is a pair of images labeled with the name of the person in the image. Some images contain more than one face. The labeled face is the one containing the central pixel of the image—other faces should be ignored as "background".

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

How many instances of each type are there?

The dataset consists of 13,233 face images in total of 5749 unique individuals. 1680 of these subjects have two or more images and 4069 have single ones.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format. Each image is accompanied by a label indicating the name of the person in the image. While subpopulation data was not available at the initial release of the dataset, a subsequent paper reports the distribution of images by age, race and gender. Table 2 lists these results.

Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there restrictions, fees or rights associated with any of the data?

Everything is included in the dataset.

Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)

The dataset comes with specified train/test splits such that none of the people in the training split are in the test split and vice versa. The data is split into two views, View 1 and View 2. View 1 consists of a training subset (pairsDevTrain.txt) with 1100 pairs of matched and 1100 pairs of mismatched images, and a test subset (pairsDevTest.txt) with 500 pairs of matched and mismatched images. Practitioners can train an algorithm on the training set and test on the test set, repeating as often as necessary. Final performance results should be reported on View 2 which consists of 10 subsets of the dataset. View 2 should only be used to test the performance of the final model. We recommend reporting performance on View 2 by using leave-one-out cross validation, performing 10 experiments. That is, in each experiment, 9 subsets should be used as a training set and the 10^{th} subset should be used for testing. At a minimum, we recommend reporting the **estimated mean accuracy**, $\hat{\mu}$ and the **standard error of the mean**, S_E, for View 2.

$\hat{\mu}$ is given by:

$$\hat{\mu} = \frac{\sum_{i=1}^{10} p_i}{10} \quad (1)$$

where p_i is the percentage of correct classifications on View 2 using subset i for testing. S_E is given as:

$$S_E = \frac{\hat{\sigma}}{\sqrt{10}} \quad (2)$$

Where $\hat{\sigma}$ is the estimate of the standard deviation, given by:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{10}(p_i - \hat{\mu})^2}{9}} \quad (3)$$

The multiple-view approach is used instead of a traditional train/validation/test split in order to maximize the amount of data available for training and testing.

图 4

原始论文提供的 Datasheets for Datasets 示例（部分）

研究者指出，创建数据表的过程并非强调自动化。这与 AI 事实清单和模型卡都有所区别，后两者致力于提高文档生成的自动化程度，以提高可解释的效率和精准度。但是 Datasheets for Datasets 的研究者认为，尽管自动化文档的编制过程很方便，但是它在某种程度上违背了设计的初衷，即鼓励数据集创建者反思创建、分发和维护数据集的过程，进而根据他们的反思来改变或优化这个过程。

在论文中，研究者叙述了开发一个 Datasheets for Datasets 的标准流程：首先制定一系列数据表的初始问题列表，然后通过创建两个示例数据表，来测试列出的问题，进而发现问题的疏漏及不足之处，并予以优化。此后，将优化过的数据表分发给两个产品团队，帮助他们基于自己的数据集创建数据表。创建完毕后，通过社交媒体等渠道公开发布，结合来自相关从业者、研究人员、外部顾问的反馈意见和评论，对数据表进行修正，比如优化问题的内容，删除重复性问题，并对问题进行重新排序以更好匹配数据集生命周期的各个阶段。

在论文中，作者希望，Datasheets for Datasets 将促进数据集的创建者和使用者之间更好的沟通，并且鼓励机器学习领域优先考虑透明度和问责制，以此减轻机器学习系统中不必要的偏见，促成机器学习结果更大的可重复性，并且有助于研究人员和其他实践者选择更合适的数据集。此外，作者还认为，数据集数据表这种形式，很大程度上能够为没有领域专业知识的其他人员提供基本信息，并有助于缓解数据集滥用的相关问题。

自 Datasheets for Datasets 的原始论文于 2018 年发布以来，数据集数据表在众多领域都受到关注。微软、谷歌和 IBM 等公司都开始在内部试行数据集数据库的形式。谷歌除了发布记录机器学习信息的模型卡片，还发布了“数据卡”(类似于数据表的轻量版本)，来呈现图像数据集的相关特征；IBM 的 AI 事实清单中，事实上也包含数据集的随附数据表。

当然，Datasheets for Datasets 还需要进一步完善。比如，它并没有提供完整的方案，来解决数据集中可能存在的偏见或风险等问题。在对与人相关的数据集创建数据表时，可能出于社会、历史、地理、文化等某些原因，会出现偏见问题，也会涉及个人数据隐私保护的问题，因此作者建议，这种情况下需要数据表的创建者与其他领域专家展开合作，衡量以何种方式更好地收集数据表的相关信息，避免偏见并且尊重个人隐私。

（四）其他可解释性 AI 工具

限于篇幅，这份报告不会对现有的所有可解释性工具进行详细说明，而只是对其中的部分常用工具进行粗略的介绍。

年份	可解释性工具
2016	ELI5
2017	Skater，Explanation Explorer，AllenNLP Interpret，TensorBoard
2019	AIX360，ACE，Captum
2020	alibi，InterpretML，LIT

图 5 常见的可解释性工具

（1）IBM：AI Explainability 360

AI Explainability 360 工具包是 IBM 推出的可解释性工具，同时提供了可诠释方法和可解释方法。该工具基于 Tensorflow、pytorch、scikit-learn 等开源工具所开发，构建了一套涵盖不同维度的解释以及代理可解释性指标的全面的可解释性方案。同时，该工具还提供了有多种解释方式，例如数据与模型、直接解释与事后解释、局部解释与全局解释等。在 AI Explainability 360 的示范案例中，用户甚至可以根据自己的不同身份，来获取不同的可解释性方案。

（2）Microsoft：InterpretML

InterpretML 是微软开发的一个可解释性工具，主要解决了以下几个问题：（1）模型调试问题——为什么模型会犯这个错误？（2）特征工程问题——如何改进模型？（3）检测公平性问题——模型是否存在歧视？（4）人机合作问题——我如何理解和信任模型的决定？（5）法规遵从性问题——我的模型是否满足法律要求？（6）高风险问题——医疗保健、金融、司法领域的模型。InterpretML 同时支持了两种不同的可解释方法：可诠释模型（interpretable model）以及可解释模型（explainable model）。其中支持的可诠释方法包括：Explainable Boosting、Decision Tree、Decision Rule List、Linear/Logistic Regression。支持的可解释方法包括：

SHAP Kernel Explainer、LIME、Morris Sensitivity Analysis、Partial Dependence。

（3）Facebook：Captum

Captum 是 Facebook（现在的 Meta 平台公司）推出的基于 PyTorch 的可解释性工具。该工具包含了集成梯度、显著图、smoothgrad、vargrad 等可解释方案的实现。它可以快速被使用在由特定领域的库（如 torchvision、torchtext 等）所构建的模型上。

（4）Google：LIT

语言可诠释性工具（Language Interpretability Tool, LIT）是一款由谷歌研究团队推出的自然语言处理可解释性工具。该工具由谷歌的 Lan Tenney 和 James Wexler 主导开发，具有可视的、基于浏览器的用户界面，可供不同开发水平的从业者使用。LIT 工具的一个特色就是初学者友好，使用者只需几行代码即可添加模型和数据，并对模型进行解释，具体说来包括以下特点：局部解释（Local explanations）；聚合分析（Aggregate analysis）；反事实生成（Counterfactual generation）；并排模式（Side-by-side mode）；高度可扩展性（Highly extensible）；框架无涉（Framework-agnostic）。

（5）Google：TensorBoard

TensorBoard 是 TensorFlow 发布的一个可视化的工具包，它为机器学习实验提供了诸多的可视化的功能和根据。该可视化工具可以编辑单个数据点并查看由此造成的推断变化，或者通过部分依赖图分析数据集中各个特征与模型推断结果之间的关系，以此来实现对于黑箱模型的解释。

（6）ELI5

ELI5 是一个著名的 Python 语言的可解释性工具包，它帮助使用者们解释机器学习模型的预测结果。ELI5 可以同时支持多个常用的机器学习工具，例如 scikit-learn、Keras、XGBOOST、LightGBM、CatBoost 等。该工具可以为线性分类模型、线性回归模型中的权重与结果进行解释，显示影响决策的重要特征，凸显文本数据等等。同时也能够以可视化的方式来解释影响图像分类器决策的相应特征。

（7）Alibi

Alibi 是一个常用的机器学习可解释工具。该工具能够为分类和回归模型提供黑盒、白盒、局部和全局解释。该工具还可以构建反事实的可解释模型。

（8）AllenNLP Interpret

AllenNLP Interpret 是艾伦人工智能研究所和加州大学的研究人员共同推出自然语言处理可解释工具。该工具侧重于对两种不同类型的实例进行解释：显著性诠释和对抗性攻击诠释。AllenNLP Interpret 工具针对于自然语言处理模型有较好的可解释性。Saliency Map 的诠释策略借助于梯度方法来将相对于每一个标记的损失进行可视化。对抗性攻击（Adversarial Attacks）的诠释方法则是通过考察两种不同的对抗性攻击来解释模型的运作规律：替换词语来改变模型的预测（HotFlip）和删除词语来保持模型的预测（Input Reduction）。这二者都是在回答一个反事实的问题：如果某些词被替换或删掉，预测会有什么变化？通过这样的方式，我们可以知道模型对于哪一些特征是敏感的，这从某种角度上来说就对作为黑箱的模型进行了解释。

（五）可解释 AI 的腾讯实践

1. 腾讯优图可解释计算机视觉实践

（1）可解释人脸识别

优图从人脸质量分和识别置信度增加人脸识别的可解释性。目前的人脸识别技术通常依赖深度学习，天然存在着可解释性问题。优图从人脸质量分和识别置信度两个方面为人脸识别技术增加可解释性。前者用于评估人脸图像质量对于识别系统的友好度，从而为人脸图片的筛选提供依据；后者则通过不确定性概率分布来建模人脸特征，为识别判断提供额外的置信度信息。

在人脸质量分方面，优图构建了质量总分、质量归因等能力。质量总分综合了多种质量退化维度，代表图像对于人脸识别系统的总体友好度。SDD-FIQA 是优图自研的一种无监督的人脸质量评估方法，发表于 CVPR2021，其核心思想是：计算样本对应的同人和非同人相似度分布，根据两个分布的差异来进一步定义人脸质量。相较于业界固有的人脸质量方法，它的优势在于

引入多样本比较使得质量评估更准确、更具鲁棒性。人脸质量归因是判断一张人脸图对应角度、模糊、遮挡、光线等维度的退化程度，为质量总分给出可视化原因。

在置信度估计方面，优图提出了一种基于超球流形对人脸识别模型的识别结果进行置信度估计的方法。该方法从理论出发，在当前 SOTA 的人脸训练 loss 上进行理论改进，并在多个人脸识别数据集上均得到提升。该方法将人脸样本特征从确定向量升级为概率分布，从而获得额外刻画样本识别置信度的能力；提出适配于超球流形的 r-radius von Mises Fisher 分布建模特征，采用完备的数学理论证明所提出的框架具有更好的可解释性。置信度方案通过对样本特征的置信度进行评估，可以为评测中的高风险样本提供筛选依据，对过滤样本的可解释程度更高。

（2）场景理解 AI 的可解释实践

在基于计算机视觉的场景理解 AI 应用上，优图通过架构设计、训练集清单、备注模块等模块化工作，提高 AI 算法整体上的透明度与可解释性，进而帮助使用人员更好地理解和使用算法系统。

一是架构设计。传统的场景理解一般通过检测模型检测出来的目标判断场景类别或者通过 CNN 直接进行图像分类。此类方法缺点明显：精度低、可解释性差、兼容扩展性差。优图为了解决这些挑战，设计了一套通用 pipeline，对于不同任务，仅仅需要在不同模块配置不同的清单即可完成最终任务的识别问题。该类方法通过模块化拆解，以及定义清晰的清单配置使得算法易扩展、可解释性强，问题溯源容易，再通过 scene graph 算法创新，提升整体识别能力。当发生识别问题时可以定位到哪个模块、哪个算法出了什么问题，诸如检测不准的问题、关系预测不准的问题、属性判断不准的问题、复杂图像感兴趣子图不准的问题等等。

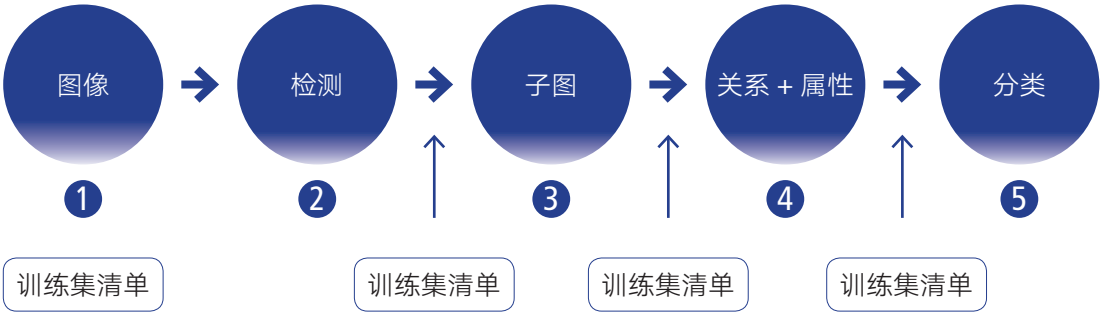

图 6 优图 pipeline 结构

二是训练集清单。介绍训练集组成情况、训练集来源等，以帮助研究人员更好地理解算法。

三是感知解析模块。通过分类、定位、分割等原子模型的一体化感知解析，提升开放世界的全场景识别能力。同时，根据细类清单配置，可以快速定位到 AI 最终决策在图像层面上的解释和依据。

四是子图构建模块。鉴于不同任务需要感知的物体的粒度不同，提供了物品超类清单（即哪些物品对于该任务没有区别），再根据超类清单合并类别，最后通过自研联通图算法将复杂图像分成几个感兴趣区域。

五是关系和属性预测模块。通过配置感兴趣的关系清单，算法可以在子图内输出对应的“物体，关系，物体”这样的三元组描述，比如“人骑电动车”“箱子在电动车上”，此外对相关的物体输出属性信息，例如“人穿着蓝色的衣服”“箱子是蓝色的”，当这一系列的关系、属性发生时，我们便有了很大的概率猜测这是某外卖。

六是分类模块。通过配置任务清单，即哪些关系、哪些属性发生时，可以判断结果，这可以直接用来分类，也可以结合分类算法融合决策。

七是备注模块。此模块主要是备注算法的基本信息包括版本、优缺点、注意事项等，帮助使用人员更好地理解和使用算法。

（3）模型安全评估

腾讯优图还提出了高效的模型安全性指标评估方法。因为训练数据有限，模型并不能“看完”所有的样本，这就导致了边界存在很多“对抗样本”。那么，如何解释 AI 模型的安全性呢？腾讯优图提出的安全性评估方法，通过输出目标 AI 模型的攻击成功率来解释该 AI 模型相对于对抗样本的鲁棒性强弱，而且无须真实的训练数据，也不需要目标 AI 模型内部的参数和梯度信息，大大降低了评估的成本和难度。具体做法就是，首先设计一个数据生成器来生成伪造的真实训练数据；然后将该数据输入到目标 AI 模型和替代 AI 模型中；接下来设计一些相似度损失来约束目标模型和替代模型输出的距离，从而达到模型窃取的目的；最后，通过攻击替代模型来实现对目标模型的迁移性攻击。

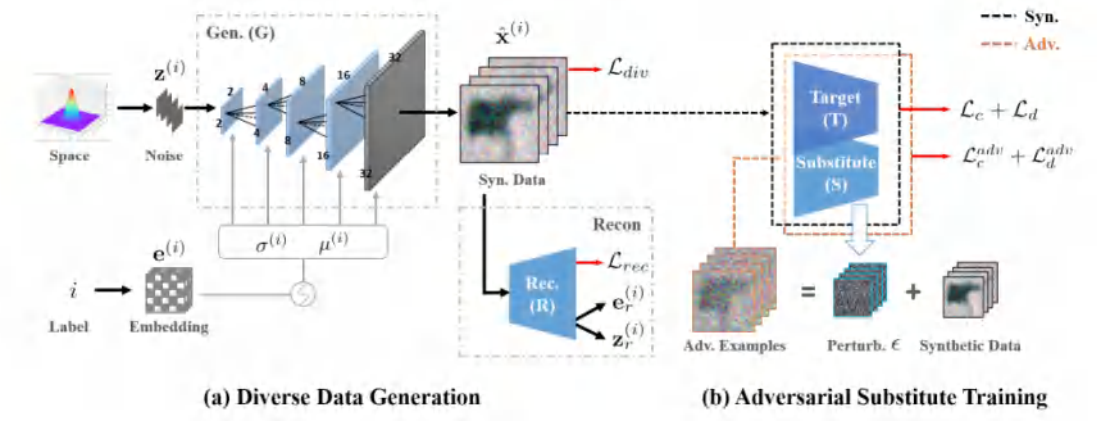

图 7 无须真实数据和被攻击目标 AI 模型内部梯度，就可以对目标模型进行安全性的评估和解释

2. 腾讯医疗 AI 的可解释性实践

（1）腾讯觅影青光眼样眼底疾病辅助诊断

临床医生与 AI 的人机协作是现阶段医疗影像分析 AI 算法落地应用的重要路径。腾讯觅影青光眼样眼底疾病辅助诊断功能，旨在通过提高医疗 AI 辅助决策模型透明度和增加模型输出参考信息多样性的方式，更好地辅助医生通过眼底影像进行青光眼样眼底疾病临床诊断。该功能通过收集构建的权威数据集训练专业的眼底影像分析模型，实现对眼底视盘和视杯区域的高精度像素级分割，以及青光眼样眼底表现的智能化识别，未来可辅助医生进行快速稳定的青光眼样眼底疾病诊断。与此同时，腾讯天衍实验室还设计了基于多任务协同的深度学习系统（如下图所示），引入视盘识别任务作为青光眼样眼底分类任务的辅助任务。具体而言，该系统在共用特征编码器的基础上，通过视盘预测任务增强对于特征提取的约束作用，使主任务的青光眼样眼底分类模型在前向推理过程中更加关注视盘区域信息。这与临床上医生进行青光眼诊断时重点关注视盘区域的机制是一致的，能有效提高青光眼样眼底分类模型输出的可解释性。

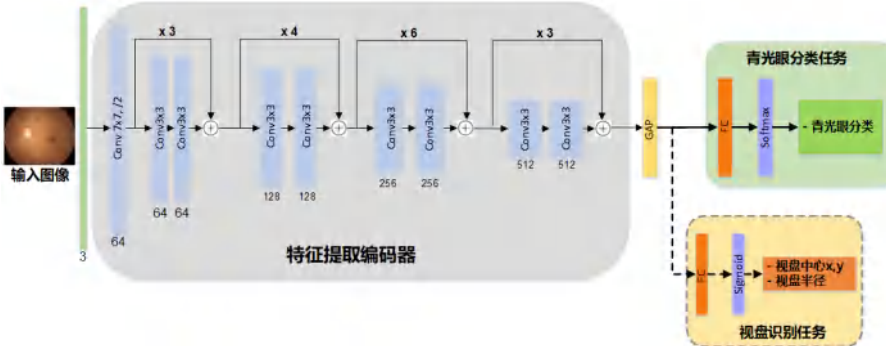

图 8
基于多任务深度学习的
青光眼样眼底分类系统

本产品是辅助诊断类医用软件，按当前医疗器械分类规则，属于第三类医疗器械，在完成研发及产品定型后，通过临床试验并取得国家药品监督管理局颁发的第三类医疗器械注册证后可正式投入临床应用。目前产品临床评价及注册工作正在开展。

在相关产品注册并实现商业化后，临床医生在借助腾讯觅影青光眼样眼底疾病辅助诊断功能进行临床诊断时，不仅能看到模型输出的“疑似青光眼样眼底表现”或者“未见明显异常”的诊断建议，还能进一步结合视盘和视杯的精确分割结果进行综合判断。医疗影像 AI 算法提供的多维度临床辅助诊断参考信息，在提升医生诊疗效率的同时，也能让医生在交互实践中逐渐增加对模型输出结果的信任度。

（2）腾讯觅影《肺炎 CT 影像辅助分诊及评估软件》

腾讯觅影《肺炎 CT 影像辅助分诊及评估软件》是腾讯首款获得国家药品监督管理局第三类医疗器械注册证的辅助诊断软件产品，用于肺部 CT 影像的显示、处理、测量和肺炎病灶识别，可辅助用于成年的新型冠状病毒肺炎疑似患者的分诊提示以及确诊患者的病情评估。供经培训合格的医师使用，不能单独用作临床诊疗决策证据。该软件的开发研发历程充分体现了腾讯对可解释 AI 的理解与实践。

首先，该软件的描述文档对用户资格进行了明确定义：经培训合格的医师。进一步地，根据产品的使用范围定义了安全性级别：用于肺部 CT 影像的显示、处理、测量和肺炎病灶识别，不给出对患者的诊断意见，不能单独作为临床诊疗决策证据，具体诊断还需要医生结合其他方式的进一步检查结果做出。本产品主要是提高医师的工作效率，使用的环境是医院发热门诊、放射科、呼吸科、感染科等需要胸部 CT 确认是否肺炎的场所，核心功能是识别肺部 CT 影像，对患者进行分类，其风险较小，不会对人体造成严重伤害或者死亡。结合法规要求和临床功能、使用环境和核心功能判断，该产品的软件安全性级别为 B 级。

该软件的使用说明书明确规定了适用的 CT 数据类型，包括 CT 设备兼容性、扫描参数要求、CT 影像质量要求等。例如对于 CT 影像质量要求，本产品算法要求输入 CT 图像为：

- 符合 DICOM 3.0 协议标准数据；
- 图像不得进行任何修改、编辑、不得进行有损压缩；

- 每例图像应保持连续完整，不得出现缺层、错层等问题；
- 肺部影像范围应涵盖肺尖到肋膈角，包含全肺；
- 图像为胸部横截面图像；
- 肺部影像不存在明显的呼吸运动伪影或其他伪影。

针对 AI 专业人员，但该软件提供了产品工作原理的详细描述，满足全局可解释性。此外，软件研究资料中对训练及测试数据的来源、数量、多维分布进行了详细分析，帮助 AI 专业人员和产品用户（如医生）理解软件背后的模型特性，消除对因训练数据偏移（bias）而导致模型输出偏移的疑虑。

特别是针对产品的目标用户（经培训合规的医生），该软件的主要功能为自动识别肺炎表征、非肺炎表征（图 9 黄色箭头），当肺炎表征中具有疑似新冠肺炎的表征时，则会提示疑似新冠肺炎。为了增强主要功能输出的可信度，并进一步支持临床量化分析，本软件在输出肺炎表征的基础上，（1）自动识别肺中的肺炎，并对肺炎的具体形状进行分割描边（图 9 绿色箭头）；（2）根据分割结果计算肺炎的体积、占全 / 左 / 右肺的大小、平均 CT 值、最大 / 小 CT 值、标准差；（3）根据分割结果分析直方图，包括左右肺直方图、肺炎直方图；（4）根据肺叶分割，判定肺炎在该患者的肺部位置（如：左肺上叶、左肺下叶、右肺上叶、右肺中叶、右肺下叶）；（5）系统会自动以影像报告的形式显示 AI 量化分析肺炎的具体信息。因此，本软件同样满足局部可解释性。

（3）腾讯天衍可信可解释的疾病风险预测模型

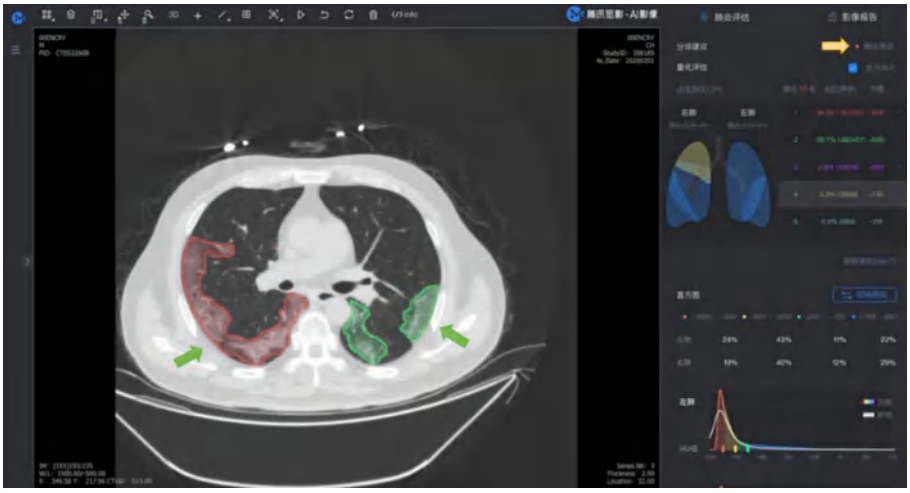

图 9
《肺炎 CT 影像辅助分诊及评估软件》界面截图。黄色箭头：主要输出（肺炎表征）；绿色箭头：增强主要输出可解释性的辅助输出（肺炎区域分割描边）

建立基于历史电子病例（electronic health record，简称 EHR）的个性化医疗预测模型已成为一个活跃的研究领域。得益于强大的特征提取能力，深度学习（DL）方法在许多临床预测任务中取得了良好的性能。然而，由于缺乏可解释性和可信度，很难将 DL 应用于实际的临床决策案例中。可解释性要求临床决策模型能够提供决策的依据，即决策的规则、参考因素以及各类因素对最终决策的权重占比；可信度要求模型能够提供 DL 模型在认知层面上的不确定度，即能够识别训练数据分布以外的样本。基于此，天衍实验室的研究人员在自解释深度模型基础上（如图 10 所示），将模型扩展为参数不确定的贝叶斯神经网络。

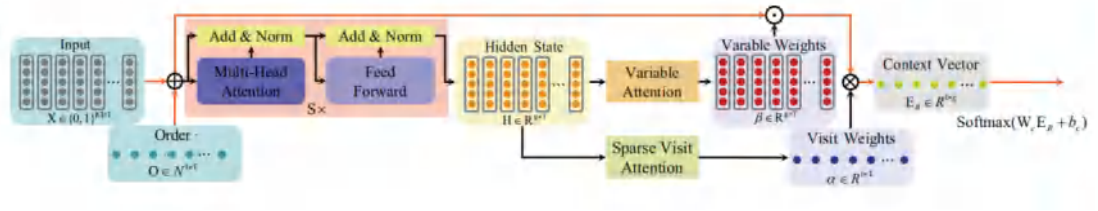

图 10 基于自解释深度学习框架的疾病风险预测模型

如图 11 所示，以心力衰竭风险评估为例，医生在借助天衍疾病风险预测模型诊断过程中，既能看到输入的各个症状对最终预测结果的贡献度（图 11 左），又能通过观察模型对该患者患病风险评估的概率分布（图 11 右，方差越大越不确定）来判断模型对该评估过程在认知层面上的确定程度。举例来说，如果所预测的概率分布方差超过一定阈值，就表明模型对该次诊断是极度不确定的。在这种情况下，天衍疾病风险预测系统会提醒医生该次诊断结果的参考价值不高，并推荐对该病例进行多专家会诊。腾讯天衍疾病风险预测模型着眼于与模型落地相关问题中极为重要的两方面，即可解释性和可信度，以此来增加模型的透明性。一方面告诉医生模型对预测结果的判断依据，另一方面能在模型自身认知不足的情况下主动提醒医生而不是不计后果地“随机猜测”。

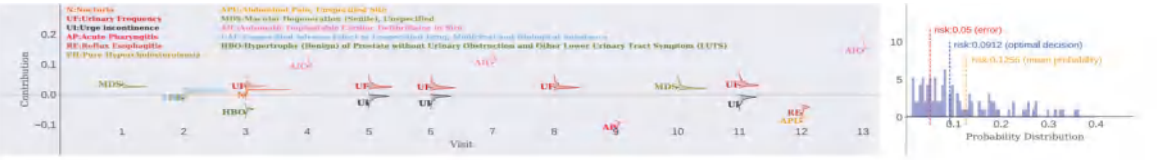

图 11 案例分析

（4）新冠肺炎症状预警检测系统

2020 年以来，新型冠状病毒肺炎疫情突如其来并席卷全球，成为人类历史上致死人数最多的流行病之一。为了提高新冠肺炎疫情检测能力和水平，落实对新冠肺炎病毒感染者“早发现、早报告、早隔离、早治疗”的要求，腾讯天衍实验室承接了深圳市疾病预防控制中心新冠肺炎症状预警检测系统建设需求，该系统通过实时分析深圳市 74 家医院的相关电子病历信息，以自然语言处理技术解析现病史、检验单、报告结论等文本信息，从中提取特征并构建新冠高风险患者识别模型，以及时提醒医院、疫情防控指挥部、三人小组关注和追踪高危疑似病例，防止疫情暴发与扩散。

图 12

疾控中心不仅要求系统对新冠高风险人群的判断具有高准确率，还关注模型输出风险评分的整体流程及其原理。基于此，系统采用了医学顶刊文章中常见的“特征抽取 + AI 分类 + 可解释性分析”的疾病预测 Pipeline，而没有选择基于文本的端到端预测模型。如图 12 所示，系统整体上分为“特征抽取”与“风险预测”两大模块。特征抽取模块首先从诊疗方案、医学文献等多个来源筛选所需特征列表，并将其组织成具有上下位关系和同义关系的本体；进而基于病历结构化、术语标准化等人工智能技术，从电子病历中提取症状、检验、影像等信息并将其映射到预先构建好的特征本体上。使用提取好的患者特征向量，风险预测模块分别从专家打分、有监督分类、相似病例匹配三条技术路线计算患者的新冠风险评分，再基于模型融合得到最终评分，并给出评分结果的可解释性分析。

从全局可解释性的角度来看，新冠肺炎症状预警检测系统将高危疑似患者识别任务拆解为多个子模块，通过查看子模块输出的中间结果，包括特征体系、原始病历、病历结构化结果、术语标准化结果、抽取的特征向量、子模型评分等，疾控工作人员不仅可以掌握系统的整体运作流程，还能对各个模块进行单独评判与分析，从而能够快速发现问题所在与系统优化的方向。

从局部可解释性的角度来看，系统主要从特征重要性和相似病例两个维度对评分结果进行解释。对于专家打分模型，用户可查看输入病例命中了哪些专家预设规则及其权重。有监督分类采用 SHAP 值或模型自带的特征重要性变量（如信息增益、回归系数）衡量特征的重要程度，通过查看输入病例的特征及其重要性，用户可理解其被判断为高风险或低风险的原因；对于相似病例匹配模型，用户通过查看与输入病例相似的历史其他病例及其诊断，在对比中可更准确地判断输入病例是否是一个真正的高危疑似患者。

四 对可解释 AI 未来发展的几点看法

透明性与可解释性，连同公平性评价、安全考虑、人类 AI 协作、责任框架，都是 AI 领域的基本问题。对 AI 应用系统的透明性与可解释性要求，需要考虑受众需求、应用场景、技术与经济可行性、时空等因素，同时与效率、安全、隐私、网络安全、知识产权保护等目的做好平衡，界定不同场景的最小可接受标准，必要时采取常规监测、人类审查等刹车机制，而非一味要求算法解释或算法公开。

（一）立法和监管宜遵循基于风险的分级分类分场景治理思路

参考国外经验，立法不宜采取过度严苛的监管要求，避免在透明度与可解释性方面对 AI 算法应用提出“一刀切”（one-size-fits-all）的要求，也不宜简单粗暴要求公开算法，而是需要采用包容审慎的立场，建立分级分类分场景的监管方式，支持 AI 企业不断创新和发展，兼顾政府、科技企业以及社会公众的整体利益，在鼓励科技创新、追求科技向善、维护社会公共利益之间找到平衡点。

具体而言，在可解释 AI 的实现方式上，首先，披露 AI 算法模型的源代码是无效的方式，不仅无助于对 AI 算法模型的理解，反倒可能威胁数据隐私、商业秘密以及技术安全；其次，不宜不加区分应用场景与时空场合地要求对所有的算法决策结果进行解释；再次，侧重应用过程中的披露义务，部署 AI 的主体对于任何披露与记录要求负有责任，需要披露 AI 系统实质性参与决策或与人互动的实施，披露应当以清晰易懂有意义，提供关于 AI 参与的关键任务的模式；最后，避免强制要求披露用来训练 AI 模型的数据集，这不仅不具有可操作性，而且容易与版权保护冲突，侵犯用户的数据隐私或违反合同义务。

（二）探索建立合理适度的、适应不同行业与应用场景的 AI 可解释性标准

整体而言，对于合理的解释，不存在一刀切的路径。具体而言，应考虑以下因素：

其一，对象。有意义的解释，可能因对象而异。普通人感兴趣或能理解的因素与复杂程度，可能与专业的审查人员或法律调查人员需要的恰当信息存在很大不同。不同的用户的需求差异很大。例如，普通用户可能想知道为什么 AI 系统作出了特定的决策，以便能够有理由质疑其决策，如果用户认为决策是不公平或错误的。因此，需要向用户提供明白易懂的，非技术语言的解释。专业人员则需要更全面的，更多技术细节的解释，以便评估 AI 系统是否满足可靠、准确等方面的一般性要求。不向普通的用户提供详尽的解释可能有悖常理，但在实践中却是有益的，这意味着向普通用户提供详尽的解释并不是必备选项。为了解释 AI 系统的预测，而给普通用户提供底层的数学公式，即便这可能是技术上最准确的解释，但普通用户却不大可能理解。因此，明白人们的真正需求才是至关重要的。普通用户也许只是希望确保 AI 系统的输入与输出是公平合理的，而非希望对背后的计算具有深层次的理解。而且即使需要对 AI 模型的功能进行彻底的理解，也不需要一次性做出。向人们提供简短的、高层次的解释，以便人们可以研究并要求额外的（必要及希望时）细节，通常是更加用户友好的方式。

其二，应用场景。应用场景的不同也可能影响提供解释的时间与方式。并非所有的应用场景都需要对 AI 算法模型及其决策结果做出事无巨细的解释，这取决于 AI 算法决策是否对受众的合法权益产生实质性的影响。例如，对于在餐厅分配位置或自动整理手机相册的算法，与审批贷款或辅助判刑的算法，需要区别对待。如果一刀切地要求提供详尽的解释，缺乏合理性和必要性。

其三，时间与地点。从目前的技术来看，要求 AI 面向全部应用场景，实时地、大规模地提供解释颇具挑战性且难以实现。行业中的可解释 AI 实践更多聚焦于不同应用场景下的事后解释。

其四，解释的关联性（目的）：为什么需要进行解释？AI 系统的目的与应用场景至关重要。相比于执行影响较小的任务的 AI 系统，如推荐电影的 AI 系统，AI 系统被用来影响生命财产安全的决策，如医疗、司法等，需要更多的投入与深度的解释。例如，法国要求在疾病诊断或治疗中使用 AI 系统，医生应当告知患者任何有助于疾病预防、诊断、治疗的大数据处理操作的存在和方式，任何对算法参数的修改都必须在卫生专业人员的参与下才能完成；任何 AI 系统都必须确保算法决策和基础数据的可追溯性，并保证卫生专业人员对这些数据的访问。

其五，技术与经济可行性。一些先进的、复杂的 AI 系统在向人类解释其运作方面可能存在技术限制。在经济可行性上，也需要考虑成本维度，大规模地提供解释所需成本与投入也需要考虑在内，以避免不合理的细节或严格的要求阻碍有价值 AI 系统的部署。尽管投入足够的时间、

精力、专业知识与正确的工具，通常可以知晓复杂 AI 系统是如何运作的，理解 AI 系统的行为背后的原因，但如果要在实践中不加区分地要求解释，不仅在规模应用上欠缺经济可行性，而且可能适得其反地阻碍具有巨大价值的 AI 系统（例如拯救生命）的应用部署。因为解释的成本十分高昂，所投入的技术资源也更加巨大。如果 AI 系统的每一个结果都强制要求完全可追溯（traceable），提供详尽的解释，这是一个相当高的标准了，那么这在实践中可能极大地将 AI 系统限制在最基本的技术（如静态的决策树）。这最终会极大地限制 AI 的社会与经济效益。比如，一个医疗算法，如果每次诊断结果都要求提供详尽的解释，可能这个算法永远无法投入使用，产生价值。因为每次输出一个决策，可能得花费数天时间来提供解释。

因此需要采取折中路径，考虑技术限制与不同可解释标准需要的利益权衡，以便平衡使用复杂 AI 系统带来的好处与不同的可解释性标准带来的实践限制。用户友好型的解释应当是准确的、清晰的、明确的、场景敏感型的、有效的，以提高对 AI 系统的整体理解：

- 解释是否准确传递了支撑 AI 系统的推荐的关键信息（key information）？
- 解释是否有助于对 AI 系统整体功能的理解？
- 解释是否清晰（clear）、明确（specific）、相关（relatable）、可执行（actionable）？
- 解释是否适当考虑了敏感性（sensitivity）？例如用户的敏感信息。
- 解释是否具有对应的可靠性度量？具体做法是引入技术手段，在构建可解释性结果的同时，输出结果的置信度，当置信度不高时，引入人类参与到可解释的过程。[23]

具体可以从以下方面来推进 AI 可解释性标准：

第一，针对 AI 系统的每一个应用场景都提供可解释性标准的指南，是不现实的，但可以针对一些示范性的应用场景提供可解释标准的指南。这能够给行业和企业提供有益参考，来平衡不同 AI 模型的性能与不同标准的可解释性要求。

第二，对于政策相关方而言，发布 AI 可解释的最佳实践做法案例集，以及具有负面影响的负

23 https://openreview.net/pdf?id=YUEFlzlG_0c

面做法，都是值得尝试的。包括用以提供解释的有效的用户界面，面向专家和审计人员的记录机制（例如详细的性能特征，潜在用途，系统局限性等）。

第三，可以创建一个说明不同级别的可解释性的图谱。这个图谱可被用来给不同行业与应用场景提供最小可接受的衡量标准。例如，如果某个失误的潜在不利影响是非常微小的，那么可解释性则不怎么重要。相反，如果某个失误是危及生命财产安全的，则可解释性变得至关重要。类似地，如果用户可以容易地摆脱自动化决策，则对深入理解 AI 系统的需求就不那么旺盛。

（三）探索可解释的替代性机制，形成对 AI 算法的有效约束

虽然可解释性是完善 AI 技术的最优解之一，但并非所有的 AI 系统及其决策都可以解释。当 AI 系统过于复杂，导致难以满足可解释性要求，或是导致解释机制失灵、效果不乐观时，就要积极转变规制的思路，探索更多元化、实用化的技术路径。目前在技术上主张的是采取适当的替代性机制，如第三方反馈、申诉、常规监测、审计等，这些替代性机制可以对 AI 算法的决策起到监督和保障作用。

一是第三方标记反馈（flagging facility）。第三方标记反馈机制允许人们针对 AI 系统提供使用上的反馈，常见的标记技术包括用户反馈渠道（“点击反馈”按钮），漏洞奖励机制等，类似于机器学习的强化学习算法，也即为用户设置反馈渠道，能够方便用户针对 AI 系统提供评价，从而形成一种有效的外部反馈与监督。用户反馈机制的优势在于允许用户分享其经历与感知，让用户的意见被听到，被认可。如果用户认为自己的反馈被重视且采取了相应的实践行动，可以在 AI 和用户之间形成正向循环，长此以往，可以增进用户对 AI 系统的信任。

二是用户申诉机制与人类审查介入。对于影响用户重大权益的 AI 系统，提供者需要向用户提供申诉的渠道，以便可以引入人类审查，保障用户的合法权益。从用户角度而言，申诉实际上是用户对 AI 的负面评价，也是重启 AI 解释的一项基本程序。通过引入用户申诉和第三方审查机制，能够对 AI 系统及其开发者形成有效监督，也是实现 AI 可责性的重要保障。比如，我国的《信息安全技术个人信息安全规范》《网络安全标准实践指南》等标准都对用户的投诉、质疑、反馈以及人工复核等机制作出了具体规定。

三是常规监测，包括严格且持续的测试、对抗测试等。红队（red team）测试作为对抗测试

（adversarial testing）的一种方式，是一种白帽黑客行为，涉及委托一个团队（内部或独立的）来尽最大努力发现系统中的问题。

四是审计机制（auditing）。审计机制作为确保 AI 可责性的重要方式，是对 AI 算法应用情况的记录、回溯和追查。审计的目标包括 AI 系统的目的，其功能与性能表现，关于模型架构的信息，训练、测试使用的数据集，内部检查，监控 AI 系统运行的组织流程，等等。通过算法审计，可以能够达到反向解释的作用，倒逼研发者在设计算法时采用更加谨慎、安全的路径，降低算法黑箱的不良影响。在未来甚至可以考虑引入区块链的机制，将对应的代码以及结果、算法过程记录到联盟链上，方便后续的审计。在部分场景，甚至可以引入形式化验证的方法，自动审查相应的合规性要求。

（四）引导、支持行业加强可解释 AI 研究与落地，确保科技向善

根据美国科技行业的经验，可解释 AI 的工作应主要由企业与行业主导而非由政府强制监管，采取自愿性机制而非强制性认证。因为市场力量（market force）会激励可解释性与可复制性，会驱动可解释 AI 的发展进步。一方面，从市场竞争的角度看，为了获得竞争优势，企业会主动提高其 AI 相关产品服务的可解释程度，从而让更多人愿意采纳或使用其 AI 相关产品服务，进而维持自身的市场竞争力；另一方面，从用户的角度看，用户会用脚投票，即如果用户不理解 AI 系统的运作，在使用 AI 相关产品服务时可能存在顾虑，这意味着可解释性不足的 AI 相关产品服务将无法获得用户的持久信任，因而用户对此类 AI 的需求也会降低。

为了争得用户和公众的信任，各大互联网平台都在逐渐揭开自己算法推荐的“黑箱”，比如在 2021 年今日头条、美团、微博等平台公司陆续主动公开其相关算法系统背后的原理。除了向公众在一定程度上透露算法原理，目前看，AI 的可解释性研究与应用也是国内外主流科技公司的主要方向之一。未来，可解释 AI 的发展仍将主要依靠企业与行业的实践和探索，而非政府的强制性监管，以适应 AI 技术的快速发展迭代。

（五）增强社会公众的算法素养，探索人机协同的智能范式

当前，人脸识别、语音识别、用户画像、自动驾驶等智能技术已风靡全球，AI 正迎来前所未有的应用热潮。可以预见的是，随着 AI 技术在各个领域的深度融合，未来人类的生产生活场

景都会逐渐离不开 AI 的应用，高度智能化的人机协同将成为新的范式。在此趋势下，AI 对于人类生活的影响程度也会日益加深。人们要想用好 AI，使之为自身创造价值，就必须走近 AI，理解 AI。一方面，增强社会公众对 AI 使用的算法素养，是构建友好型人机协同范式的必然要求；另一方面，增强公众的算法素养，是公众防范算法侵权，维护自身利益的重要途径，能够推动可解释 AI 的规范化和实效化。针对提高公众素养的问题，欧盟立法者认为，可以通过教育、新闻报道、揭秘等方式提高公众的算法素养。当公众可以理解 AI，AI 将不再是公众眼中的洪水猛兽，而是一个具有极强创造力、执行力及忠诚度的“好伙伴”，希望自己的想法能被 AI 执行，AI 的决策能为自己理解，形成人机之间的友好交流机制。

五 结论：
寻找平衡的 AI 可解释性路径，确保科技向善

总体而言，透明性与可解释性本身不是目的，而是增进责任与问责，赋能用户，打造信任与信心的方式和手段。在设计透明性与可解释性要求时，相关主体需要考虑他们想要实现什么目标，以及在特定情境下如何更好地匹配这些目标。

第一 › 不同的主体在不同的情形下需要不同形式的透明性。透明性要求必须根据不同主体的需求量体裁衣，方便他们可以理解，而非造成不必要的困扰。在一些情形下，关于特定决策的详尽的信息也许是重要的，但是在其他情形下，提供关于 AI 系统如何开发及运行的一般信息是合适的、更有助益的。

第二 › 需要平衡透明度要求与其他重要的目的，如效率、安全、隐私、网络安全等等。例如，一些形式的透明性看似有吸引力，但却可能带来相当严重的风险，而且对于增进责任与打造信任几乎无甚助益。例如，披露源代码（source code）或单个用户的数据，无助于理解 AI 系统如何运行，以及它为何做出特定决策，但却可能让 AI 系统被滥用或操纵，给用户隐私与知识产权带来显著风险。分享、开放源代码是最低端，最无效的算法透明方式。因为开放源代码无助于理解 AI 系统，因为 AI 系统太过复杂，专家也无法测量。算法不是越透明越好，例如，把算法变得简单，可以增加可解释性，同时却可能让算法更不准确。引入人类干预可以明确主体责任，但却可能由于人类的错误而降低准确性。此外，在可解释与准确性之间，如果 AI 应用对性能要求不那么高，则可解释性可以超过准确性；如果安全是优先的，则可解释性可以让位于准确性，只要存在能够确保问责的替代性即可。正如联合国《人工智能伦理建议书》所指出的那样，公平、安全、可解释性这些原则本身是可取的，但在任何情况下这些原则之间都可能会产生矛盾，需要根据具体情况进行评估，以管控潜在的矛盾，同时考虑到相称性原则并尊重个人权利等。

第三 › 人工智能的评价标准不应是“完美级”，而应在与既有流程或人类决策对比的基础上，界定评价 AI 的最低可接受标准，追求可解释 AI 的透明、公平、安全的技术价值。所以即使 AI 需要解释，也必须考虑可解释的程度。要求 AI 系统满足可解释性的“黄金标准”（远远超过既有的非 AI 模式即人类决策所要求的），可能不当地阻碍 AI 的创新性使用。

第四 › 当设定可解释性标准时，需要考虑可操作性、务实性。很难遵从或者遵从成本很高的标准会阻碍 AI 系统的应用。如果在所有的情形下都要求最详尽的解释，而不考虑实际的需求，则可能会阻碍创新，也会给企业及社会带来高昂的经济成本。适当的可解释性标准不应超过合理且必要的限度。举例而言，社会不会要求航空公司向乘客解释为什么飞机采取了算法决定的航线。类似地，一个相似的务实性、情境特定的路径应适用于 AI 的可解释性标准要求。

最后 › 即使 AI 系统并非完全可解释，我们也可以利用 AI 系统来提高决策的透明度。对人类决策的解释，也许不能准确反映出影响人类决策的因素或无意识偏见。实际上，即使 AI 系统所做出的决策并不能被完全解释，但相比理解人类如何做出类似决策，我们也可以更好地理解 AI 系统在整体上是如何做出决策的。而且，AI 的最大价值在于可以在复杂情形中发现、识别超出人类理解的模式（pattern），因此根据定义，这样的 AI 系统不会以人类可以理解的方式具有完全的可解释性。就像取得驾照，相信汽车可以安全驾驶，并不需要人人都成为专业的汽车工程师一样，当使用 AI 系统时，解释并不总是必须的。长远来看，政府、社会、企业、行业、科研机构、用户等主体需要共同探索科学合理的可解释 AI 落地方案及相关的保障与防护机制，推动科技向善。

腾讯研究院

腾讯研究院是腾讯公司设立的智库机构，旨在依托腾讯公司多元的产品与服务，围绕互联网发展的焦点问题，以科技向善为指南，通过开放合作的研究平台，汇集各界智慧，共同推动数字社会健康、有序的发展。我们坚守开放、包容、前瞻的研究视野，致力于成为现代科技与社会人文交汇的研究平台。

腾讯天衍实验室

腾讯天衍实验室成立于 2018 年 9 月，是一个专注于医疗人工智能的实验室。我们致力于通过机器学习和大数据分析在医疗领域打造一个实时演进的知识和决策平台，通过对医疗领域数据的收集、分析和建模，为公司医疗业务及产品输出技术，加速 AI 产品在医疗领域的市场化，服务于更安全和可信赖的用户体验。我们也积极进行前沿科技探索，成立至今已经发明 400 多项医疗 AI 专利、发表 90 多篇顶会和顶刊论文、并获得 10 多项国际医疗 AI 相关竞赛的冠军。

腾讯优图实验室是腾讯的人工智能实验室之一，专注于视觉技术的研究与落地，在多项国际顶级 AI 比赛中斩获赛道冠军，拥有超过 1000 件全球 AI 专利。截止目前，腾讯优图通过腾讯云输出超过 20 项 AI 解决方案，100+AI 原子能力，为国家人口普查，健康码，粤港澳小程序等提供核心的自研 AI 技术能力，为金融、工业、汽车等传统企业提供自研的 TI-one 机器学习平台服务。此外，腾讯优图专注未来科技研究，与 50+ 全球顶级院校展开产学研合作，打造如跨年龄 AI 寻人、青少年内容审核、AI 探星等技术能力，践行公司科技向善的使命和愿景。

腾讯 AI Lab 是腾讯的企业级 AI 实验室，于 2016 年 4 月在深圳成立，目前有 100 多位顶尖研究科学家及 300 多位应用工程师。借助腾讯丰富应用场景、大数据、计算力及一流人才方面的长期积累，AI Lab 立足未来，开放合作，致力于不断提升 AI 的认知、决策与创造力，向“Make AI Everywhere”的愿景迈步。腾讯 AI Lab 强调研究与应用并重发展，基础研究关注机器学习、计算机视觉、语音技术及自然语言处理等四大方向，570 多篇研究论文已覆盖国际顶级学术会议；技术应用聚焦在社交、游戏、内容与平台 AI 四大领域，在微信、QQ 等 100 多个产品中落地；行业应用不断取得突破，研发出屡获国际大奖的围棋 AI「绝艺」，与王者荣耀联合研究的策略协作型 AI「绝悟」，推进多模态虚拟人前沿技术，支持国家级 AI+ 医疗标杆产品「腾讯觅影」与「腾讯医典」，自研智能显微镜及 AI 药物发现平台「云深」，并初步探索 AI+ 农业智慧温室项目等。详情可访问：ai.tencent.com。

研究顾问：

司　晓　腾讯研究院院长

吴文達　腾讯健康副总裁

郑冶枫　腾讯杰出科学家 | 天衍实验室负责人

吴运声　腾讯云副总裁 | 腾讯优图实验室总经理

张正友　腾讯首席科学家 | 腾讯 AI Lab 和腾讯 Robotics X 主任

研究策划：

张钦坤　腾讯研究院秘书长

周政华　腾讯研究院资深专家

写作团队：

腾讯研究院：曹建峰 | 王焕超

腾讯天衍实验室：魏东 | 黄予 | 张先礼 | 孙旭

腾讯优图实验室：丁守鸿 | 尹邦杰 | 陈超 | 黄余格

厦门大学：詹好

研究支持团队：

马锴 | 李博 | 王强 | 赵子飞 | 李南 | 刘金松 | 吴保元 | 卞亚涛 | 吴秉哲

黄俊 | 陈瑶 | 田小军 | 朱开鑫 | 胡锦浩 | 梁竹

研究联系：

腾讯研究院 曹建峰

邮箱：jeffcao@tencent.com 微信：soyjfc