智能机器

可解释人工智能：可信赖机器的关键

我们怎样才能学会信任机器？其关键在于推进可解释人工智能（XAI）的发展。ABB 回顾了 Leviathan 旗下公司进行的开创性研究，并探讨了这一业务关键领域的最新研究。

Jinendra Gugaliya
ABB 集团研究中心
过程自动化事业部
印度班加罗尔

jinendra.gugaliya@in.abb.com

K. Eric Harper
ABB 前员工

人工智能（AI）模型已渗透到我们的日常生活中，但我们往往意识不到。例如，Netflix 使用推荐引擎向用户推荐电影。这一看似简单的过程运用机器学习（ML）来帮助相关算法根据用户活动自动作出数百万个决定[1]。Facebook 使用其专利技术“计算机视觉接触检测系统”来识别物体，它可以对用户账户上的标志和品牌图像进行分类，以便广告商在赞助的故事帖子中向用户定向投放广告[2]。除了娱乐和广告外，大数据技术和 ML 也正渗透到任务关键型应用中，例如疾病检测和 诊断、贷款申请决策和自动驾驶汽车。这些应用可能会对我们的生活产生至关重要的影响。观看哪部电影可能无关痛痒，但抵押贷款是否发放却与我们密切相关。问题的症结在于：即使是数据科学家也难以解释这些 AI 模型，因此妨碍其接受。如果不能解释这些模型或不具有“可解释性”，这些模型很可能不会被接受或使用。此外，《通用数据保护条例》（GDPR）要求数据处理须更加透明，且 AI 流程须更加清晰，因此模型解释至关重要[3]。

> 即使是数据科学家也难以解释 AI 模型，因此妨碍其接受。

无论涉及人还是机器，决策原理都可以促进对模型背后动机的理解，并产生紧迫感；这将增加最终建议被接受和实施的可能性。没有这种理解，就不可能信任我们所制造的机器。因此，ABB 研究人员深入探究了可解释 AI 的历史，以阐明既往经验教训可如何促进未来 AI 的扩展。

早期工业 AI 的原理

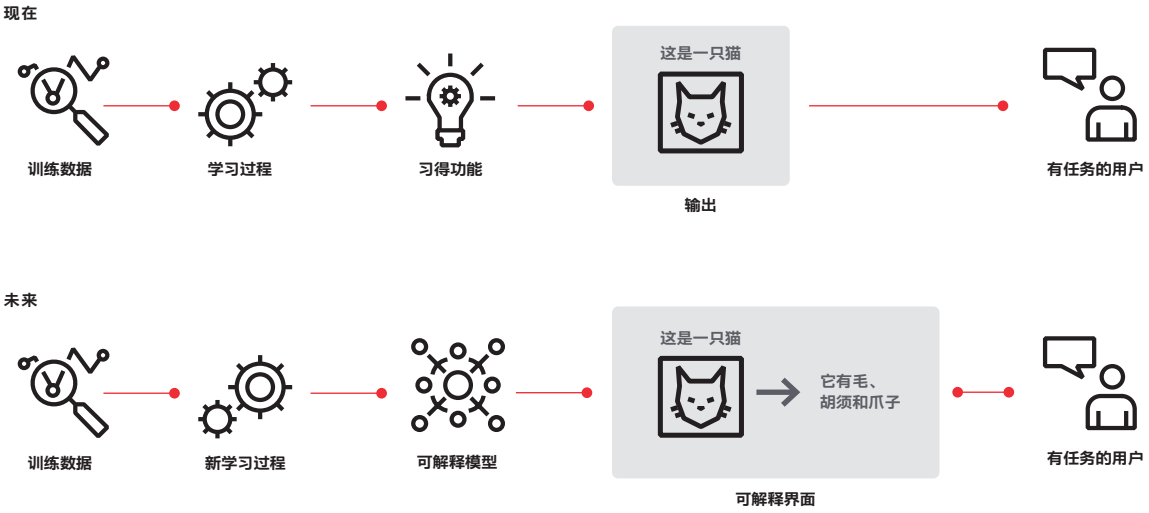
这一研究领域目前非常活跃，它始于对诊断应具备可辩证性的要求。20 世纪 80 年代，Westinghouse 与卡内基梅隆大学共同发明了首个以操作为中心的知识型决策支持系统，并将其商业化[4]。本文中提到的成就的贡献者之一 Eric Harper 直接参与了 GenAID、Turbin AID 和 ChemAID（人工智能诊断）项目，并基于一项关键专利促进了知识产权和软件技术升级，该专利的重点是一个操作系统[5]，可提供证据和最佳实践来处理异常情况。Westinghouse 的后续专利描述了一种方法和系统，可通过充分练习知识库确认预期结果来自已知的异常数据输入[6]。Harper 创建了

ABB 研究人员深入探究了可解释 AI 的历史，以阐明既往经验教训可如何促进未来 AI 的扩展。

用于探索知识和跟踪支持的工具和技术，以证明基于诊断的具体行动是合理的。这一关键知识产权现已公开，而这些发现至今仍具有现实意义[7]。

AI 结果解释方面的创新

最近，一家著名的软件咨询公司采纳了基于知识表示生成模型的想法。其在一项美国专利



01

申请中提出，通过将 k 均值聚类、主成分分析、正向或反向链接以及模糊逻辑相结合，可解决在解释 AI 结果方面实际面临的一般问题[8]。利用其他方法，管理信用风险的信用评级机构可使用不同输入来充分运行模型，有效地向客户展示各种输入如何导致不同输出，从而帮助客户理解所做出的信用决策[9]。在一项美国专利申请中，英特尔提出了一种技术，该技术可识别在机器学习训练阶段观察到的结果与在操作过程中获得的结果之间的差异[10]。在英特尔的另一项专利申请中，描述了神经网络对可解释 AI 的影响：这一过程可全面跟踪较低与较高神经网络层之间的依赖关系和支持力度，并以一种看似令人惊讶的方式追溯到其输入特征，如同前向链接专家系统一样[11]。谷歌将这些想法与其工具和框架相结合[12]；IBM 也建立了一个类似平台[13]。

可解释 AI 的趋势

在 XAI 中，这种重新兴起的研究重点可根据透明度分为不同的类别：从全黑盒（低透明度）到白盒（高透明度）特征分级[14]：

- 不透明系统
- 可理解系统
- 可解释系统

尽管这为 XAI 模型设计和开发带来诸多好处，但会大幅增加成本。因此，需要在可解释性和准确性之间 进行权衡。这种平衡和透明度将取决于业务需求，以及相关应用的实际采用方式。

ML 模型决策背后的逻辑复杂且不明显。在未经治理的情况下，信任由此作出的关键决策是有问题的。这种担忧从一开始就令人却步：神经网络在 20 世纪 80 年代已经可供用于机器诊断，但并没有付诸实施。因此，在特定领域中提供 XAI 变得越来越重要。这样，就可以对 XAI 模型进行形式验 证，而这种能力对于医疗应用尤为重要，因为医疗建议关乎生死。

XAI开发将取决于业务需求，以及模型的实际采用方式。

在 ML 模型可以被轻易接受之前，还需要解决另一挑战：如果训练数据无法覆盖整个解决方案空间，会出现偏差[15]。如果在各种条件下进行测试来衡量解决方案的优缺点，就有可能发现这种偏差缺陷。

尽管如此，XAI 模型仍能带来令人赞叹不已的新洞察。如今的 ML 系统通过 数百万个例子加以训练，因此可以识别出人类难以发现的数据模式。Westinghouse 的早

—
01 示意图说明了在现在和将来如何使用可解释 AI 打开“黑盒”。此模型由 FICO 开发[9]。

—
02 显示了对决策的实际支持，根据 Bellows 等人的文章[23]重新绘制。决策依赖于计算置信度（CF）、严重度（SEV）和重要性（IMP）。

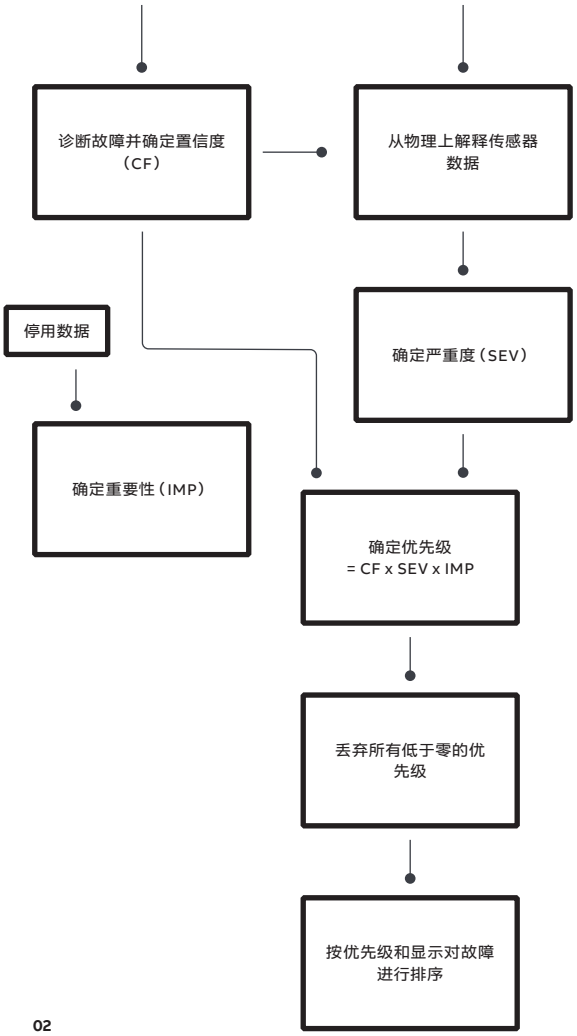
期梦想是，希望工程师们有朝一日可将从 GenAID 等模型收集的数据导出、应用数据科学方法以及探索新创新，这一梦想已基本实现。使用 XAI 系统促进了新洞察的出现：人们可以从机器学习中提取提炼出的知识来获得新的感知，例如围棋游戏 Go[1] 的新策略是用 ML 开发的，而现在人类玩家也在使用。

尽管如此，不透明的决策支持对企业并没有吸引力。显然，银行必须说明拒绝信贷申请的原因。而且，AI 模型必须符合法律要求，为所生成的决定提供证据。

这些问题的普遍化导致欧盟调整了实施“解释权”的新规定，从而使用户有权要求获得相关算法决策的原理。人们希望，XAI 能够在训练和运行模式中提供 ML 模型所需的信心、信任、公平和安全，以赢得业务机会[16]。

—
脚注

1) Go 是一款双人抽象策略棋盘游戏。



02

—
ML 系统通过数百万个例子加以训练，因此可以识别出人类难以发现的数据模式。

解释 AI 模型：现状

目前，有两种方法被用于提高 AI 模型的可解释性。第一，鉴于解释目标从本质上选择模型结构。第二，对复杂的 AI 模型进行逆向工程，使其可理解。然而，用易于理解的原理设计模型会影响准确性，反之亦然，例如：复杂的深度神经 网络（DNN）尽管准确，但缺乏可解释性。诸如线性回归或基于决策树的模型等算法更易于解释，但准确度较低。在 AI 模型的准确性与可解释性之间取得平衡是当前的研究热点[17]。

在有关 XAI 的另一研究领域中，探讨了局部与全局可解释性之间的差异[18] →03。局部视角基于敏感性分析（SA）原则确定模型输出如何随输入或调参扰动而变化。虽然不产生函数值本身的解释，但 SA 可确定解释模型结果的因素和配置。全局视角则使用两种技术。逐层相关性传播（LRP）可反向重新分配预测函数，从神经网络的输出层开始，然后反向传播到输入层。LRP 通过分解来解释分类器的决策，可以用热图来表示[19]。而数据驱动型入侵检测系统（IDS）是一种对抗性方法，用于查找所需的（输入特征的）最小修正，以正确分类一组指定的错误分类样本。修正量可将解释错误分类原因的最相关特征可视化。研究人员曾将 LRP 和 IDS 组合用于参与由深度强化学习（DRL）驱动的 Atari 游戏[20]。

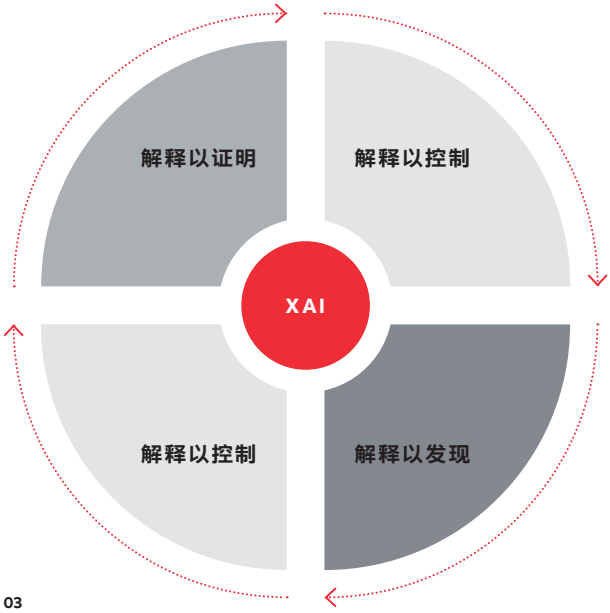
XAI 领域的研究已经扩展 到可以进行数据集比较。在设计一种可以对各种数据集[21]进行数值比较的系统时，同时使用语言原型摘要和模糊规则，并使用自然语言解释差异。所谓的 人机协同是 XAI 研究中的一个关键因素，其中可用性是模型设计和使

用的重要考虑因素之一。因此，ML 模型应该允许用户基于迭代学习以交互方式调整模型[22]。

ABB 整合 AI 先进技术来改善预测性维护、优化和性能，从而为客户创造价值。

实用的决策支持方法

工程学告诉我们，任何复杂的问题都可以通过以下步骤解决：将工作分解成多个部分，分别进行构建和验证，然后再重新组合到一起，如此形成一套完整的解决方案。20 世纪 80 年代，一种用于解释性决策支持和资产管理的 Westinghouse 系统发明了一些技术[23]，可先后通过验证测量值和将当前条件融合到系统诊断和建议中来确定工厂设备维修的优先顺序。一家大型电气



03

04

设备制造商持续使用该系统对发电厂和涡轮发电机进行现场监控，并在其电力诊断中心运行该服务[24]。基本上，潜在问题可以根据以下三个维度进行排序：

- CF——诊断置信度
- SEV——失效倒计时
- IMP——故障和修复最严重损坏产生的成本

这些维度是针对导致故障或停运的每个设备传感器、部件和系统计算的：使用这一组合来确定针对每种可能故障的行动优先级。排序原理是可向客户解释的：通过部件和传感器读数来追溯三大维度和诊断细节。如此，客户便可以直接接受排序，而无需深入了解细节。

与 ABB 的相关性

在 20 世纪 80 年代，由于计算能力非常弱，因此对自动化服务解决方案进行了调整，以利用有限的资源提取出最佳性能。如今，即使对于部件的最细小颗粒，这些最初解释特征中的每一个都可以计算出来，然后根据它们之间的关系和依赖性进行组装。ABB 擅长利用从早期工作中积累的知识和流程建立其自动化服务解决方案。如今，ABB 依赖大量状态监测解

05

06

—
03 可解释 AI 让用户能够理解并接受决策。发现、控制和证明的能力都是相互关联的。

—
04 在 ABB，AI 具有巨大潜力，因此被用于众多工业领域。例如，在工厂的工业分析过程中，工程师可以标记感兴趣的模式，然后用来训练基于 RNN 的分类器。

—
05 通过在 ABB 传感器诊断中设计可解释 AI，使工厂操作和管理人员能够理解并接受需要执行的工作。

—
06 客户得益于可解释 AI，无需理解用于决策的所有计算。

决方案，无论何时，只要有问题的设备产生麻烦，这些解决方案就会指示置信度→04–06。 ABB Ability™ 高级数字化服务解决方案包含解决失效前时间这一关键问题的功能。因为拥有优秀的工程师和信息科学家，所以 ABB 能够为计算如下重要维度制定数据科学解决方案：CF、SEV 和 IMP（如适用）。对故障后维修总成本的深刻理解使 ABB 能够根据 数据对问题进行诊断；这催生了诸如 ABB Ability™ 等新服务。

通过在应用程序中设计 XAI，ABB 在市场上脱颖而出：这可以促进信任 – 其重要性更胜以往。

作为工业自动化领域的先行者，ABB 通过整合 AI 先进技术 来改善预测性维护、优化和性能，从而为客户创造价值。尽管如此，工厂操作员和管理人员在实施成本可能高昂的行动之前，必须了解由运营和服务应用程序推荐的 AI 模型所生成决策背后的理由和原理→05。通过在应用程序中设计可解释 AI，ABB 在市场上脱颖而出：这可以促进信任 – 其重要性更胜以往。当模型可解释时，专家和最终用户便可以确信结果是无偏见、安全、合法、道德且适当的。 •

参考文献

[1] Code Academy, “Netflix Recommendation Engine”, 来源: https://www.codeca- demy.com/.[访问日期: 2020 年 5 月 5 日].

[2] A. Razaaq, “Facebook’s New Image Recognition Algorithm Can Scan your Picture for Advertising Opportunities” in *B2C Business to Community*, May 19, 2019, Available: https://www.business-2community.com/.[访问日期: 2020 年 5 月 5 日]

[3] A. Woodie, “Opening up Black Boxes with Explainable AI” in *Datanami*, May 30,2018, 来源: https:// www.datanami.com/.[访问日期: 2020 年 5 月 5 日].

[4] E.D. Thompson, et al., “Process Diagnosis System (PDS) – A 30 Year History” in *Proc. 27th Conf. on Innovative Applications of AI*, Jan. 2015, 来源: https://dl.acm.org/doi/10.5555/2888116.2888260. [访问日期: 2020 年 5 月 5 日].

[5] Thompson, et al., “Methods and appa- ratus for system fault diagnosis and control”, US Patent no. 4,649,515, March 10, 1987.

[6] K.E. Harper, et al., “Expert system tester”, US Patent no. 5,164,912, November 17, 1992.

[7] Y. Lizar, et al., “Implementation of Computer Damage Diag- nosis by Expert System Based Using Forward Chaining and Certainty Factor Methods” in *International Journal of Scientific & Technology Research*, vol. 8 issue 6, June 2019, pp. 141–144. 来源: http://www.ijstr.org/. [访问日期: 2020 年 5 月 5 日].

[8] L. Chung-Sheng, et al., “Explainable Artificial Intelligence”, US Patent Application no. 20190244122, August 8, 2019.

[9] FICO, “How to Make Artificial Intelligence Explainable: A new Analytic Workbench”, in *FICO/blog*, Sept. 13, 2018, 来源: https://www.fico.com/blogs/. [访问日期: 2020 年 5 月 5 日].

[10] J. Glen, et al., “Misuse Index for Explainable Artificial Intelligence in Comput- ing Environments”, US Patent Application no. 20190197357, June 27, 2019.

[11] K. Doshi, “Mapping and Quantification of Influence of Neural Network Features for Explainable Artificial Intelligence” US Patent Application no. 20190164057, May 30, 2019.

[12] GoogleCloud, “Understand AI Output and Build Trust”, 来源: Google Cloud: https://cloud.google.com/. [访问日期: 2020 年 5 月 5 日].

[13] A. Mojsilovic, “Intro- ducing AI Explainability 360”, August 8, 2019, 来源: IBM Research Blogs: https://www.ibm.com/blogs/. [访问日期: 2020 年 5 月 5 日].

[14] D. Doran, et al., “What does explainable AI really mean? A new Conceptualization of Perspectives”, in *AriXiv*, October 2, 2017, 来源: https://arxiv.org/abs/1710.00794. [访问日期: 2020 年 5 月 5 日].

[15] N. Mehrabi, et al., “A Survey on Bias and Fairness in Machine Learning” pre-print based on work supported by DARPA, in *ariXiv*, September 17, 2019, 来源: https://arxiv.org/abs/1908.09635. [访问日期: 2020 年 5 月 5 日].

[16] M. Miron, “Inter- pretability in AI and its relation to fairness, transparency, reliability and trust” in *European Commission WITH HUMAINT*, September 4, 2018, 来源: https://ec.europa.eu/jrc/communities/en/com- munity/humaint. [访问日期: 2020 年 5 月 5 日].

[17] W.J. Murdoch, et al., “Definitions, methods, and applications in interpretable machine learning”, in *Proc. of the National Academy of Sciences of the United States of America*, vol. 116, issue 44, 22071 October 29, 2019, 来源: https://www.pnas.org/.[访问日期: 2020 年 5 月 5 日].

[18] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)” in *IEEE Access*, September 17, 2018, 来源: https://ieeexplore.ieee.org/document/8466590. [访问日期: 2020 年 5 月 5 日].

[19] W. Samek, et al., “Evaluating the Visualization of What a Deep Neural Network Has Learned” in *IEEE Transactions. On Neural Networks and Learning Systems*, Nov. 2017. [Abstract]. 来源: https://ieeexplore.ieee.org/abstract/docu- ment/7552539. [访问日期: 2020 年 5 月 5 日].

[20] H. Jo, and K. Kim, “Visualization of Deep Reinforcement Learning using Grad-CAM: How AI Plays Atari Games?” in *IEEE Conf. on Games*, August 23, 2019, 来源: https://ieeexplore.ieee.org/document/8847950

[21] A. Jain, et al., “Explainable AI for Dataset Comparison” in *IEEE Int. Conf. on Fuzzy Systems*, June 26, 2019, 来源: https://ieeexplore.ieee.org/document/8858911. [访问日期: 2020 年 5 月 5 日].

[22] A. Kirsch, “Explain to whom? Putting the User in the Center of Explainable AI” in *Proc. 1st Int. Workshop on Comprehensibility and Explanation in AI and ML*, October 21, 2018, 来源: https://hal.archives-ouvertes.fr/hal-01845135/ [访问日期: May 5, 2020].

[23] Bellows, et al., “Automated system to prioritize repair of plant equipment”, US Patent no. 5,132,920, July 21, 1992.

[24] I. Becerra-Fernan- dez and R. Sabherwa, “Knowledge Application Systems: Systems that Utilize Knowledge” in *Knowledge Manage- ment: Systems and Processes*, 2nd ed. New York: Routlege, 2015, pp. 3–105.