并非所有解释方法都生而平等：人工智能可解释性的分层评估

NICOLAS SANCHEZ 和 BIN YU✉，美国加利福尼亚大学伯克利分校

随着人工智能过渡到社会技术部署阶段，不可解释模型对可靠性、公众信任和可争辩性构成了严重威胁。然而，并非所有解释方法都生而平等：要使可解释性提升部署安全性和有效性，就必须在分层体系中评估解释方法，以区分不同的可信度证据等级，而不是通过孤立的解释实例进行验证。

我们的核心愿景是，将解释方法的评估提升为人工智能评估指标的核心组成部分，从而推动稳健的数据知识发现（KDD），并确保人工智能社会技术部署的安全性与可争辩性。为此，我们提出TRUST——一种分层评估框架，其建立在真实数据科学（VDS）的可预测性、可计算性与稳定性（PCS）原则 [22, 30, 31, 33]以及预测性、描述性与相关性（PDR） [19] 基础之上。为了有效且系统地区分解释方法，TRUST通过五个层级对解释进行分级：任务对齐的 PDR 基础、严格的认识论基础、统一语义理解、范围可迁移性，以及针对性前瞻性验证。我们展示了在 TRUST 框架下对迭代随机森林（iRF）系列进行的端到端案例研究，并将其与其他难以通过特定层级的大语言模型解释方法进行对比。最后，我们简要总结了关于 TRUST 原则如何扩展至自主智能体框架的初步思考。归根结底，TRUST框架将可争辩性和认识论基础确立为人机工作流以及最终自主智能体工作流的基础性评估原则。

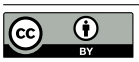
CCS概念： • **计算方法 → 机器学习；人工智能；** • **人 - 居中的计算 → 人机交互 (HCI)**。

附加关键词和短语：可解释机器学习、评估框架、TRUST 框架、真实数据科学、可争辩性、可预测性-可计算性-稳定性 (PCS)、人工智能治理

1 引言

人工智能（AI）模型正迅速从封闭循环的新奇产物转变为嵌入式社会技术系统。从英国国家医疗服务体系英格兰分部为超过五十万名临床医生提供人工智能工具 [21], 美国人口普查数据显示，37%的大型企业和55%的劳动者正在使用人工智能 [20, 26], 到不断扩大的课堂合作 [13], 人工智能已深度融入日常工作流。然而，快速采用也引发了由用户自满和自动化偏见驱动的、意料之外却十分严重的脆弱性：临床医生否定了自己准确的医学判断 [24], 律师在不知情的情况下提交了捏造的判例 [25], 教育工作者则错误地指控非母语学生存在不当行为 [17]。这些失效已引发重大反应，包括挪威禁止小学使用人工智能 [2], 加利福尼亚州建立州级人工智能安全验证框架[10], 以及国会调查 UnitedHealth 的 nH Predict 不当拒绝提供医疗服务的问题 [8]。归根结底，这些例子揭示了一个系统性缺陷：当前的评估框架依赖脱离情境的基准测试，忽视了部署现实。尽管绿色防护等范式[16] 会在日常变化下评估输出，并提出缓解策略以应对失效模式，但人工智能系统必须更进一步，满足可争辩性这一根本性的社会原则[27]。

作者联系方式：Nicolas Sanchez，nicolas_sanchez_ep@berkeley.edu；Bin Yu，binyu@berkeley.edu，加利福尼亚大学伯克利分校，美国加利福尼亚州伯克利。



本作品采用知识共享署名 4.0 国际许可协议授权。

arXiv:2611.31833v1 [cs.AI] 24 Nov 2026

可争辩性历来指的是这样一套机制：用户能够理解、构建、塑造并质疑模型预测或决策 [18]。这些机制确立了问责性，并赋予社会部署系统以可引导性。Bamberger、Mulligan 和 Wang [4] 已证明，有必要将这一概念扩展到事后特征之外：可争辩性必须是一种系统性属性。人工智能系统必须具备反思性架构，旨在揭示其中嵌入的价值选择，并提供结构化路径，以便用户不仅能在输出层面，还能在配置和设计阶段质疑模型推理。可信人工智能的模型可解释性为满足这些要求提供了不可或缺的途径，同时还能指导可信人工智能模型的开发，并确保其能够从数据中进行严谨的知识发现。然而，并非所有人工智能解释方法都同等有效。脆弱且不可靠的解释方法会适得其反，既不利于知识发现，也不利于可争辩性：它们对人工智能模型推理提供了错误认知，使虚假的安全感进一步加深了社会对人工智能系统的不信任。

为了更好地评估满足可信部署和可争辩性要求的可信解释方法，我们提出了一个五级评估框架，用于根据忠实度、安全性和效用对解释方法进行分级。每个层级都要求逐步增强的独立证据，从基线要求一直到黄金标准的外部验证，具体包括：任务对齐的预测性、描述性和相关性（PDR）基础；严格的认识论基础；统一语义理解；范围可迁移性；以及针对性前瞻性验证（TRUST）。除提出该框架外，我们还提供了一份标准化的 TRUST卡片[1]，既用于记录评估结果，也用于指导开发。TRUST 层级建立在两大既有支柱之上，旨在实现可信且安全的人工智能，包括可争辩性。首先，它直接拓展了 Murdoch 等人 [19]提出的可解释机器学习基础定义。其次，它遵循真实数据科学（VDS） [22, 30, 31, 33]的原则。

在真实数据科学（VDS）中，可预测性、可计算性和稳定性（PCS）的核心原则会在人工智能流水线的每个阶段审视合理的人类判断，并汇总其中良好或经过现实检验的判断 [33]。TRUST框架以PCS为基础，评估贯穿人工智能生命周期各阶段的解释方法的可信度——从初始设计、训练和部署前验证，到在线监控和事后审计。为满足具有不同领域和技术专业知识的利益相关者的需求，TRUST提供了一个多维度的分层评估框架，后续层级要求逐步更强的外部独立证据。

在本文余下部分，我们将详细介绍TRUST的五个层级，随后使用生物学场景中的迭代随机森林（iRF）以及其他大语言模型解释方法（CD-T、SPEX、GCT）进行从头到尾的演示。在总结性讨论中，我们将分享有关将TRUST扩展至自主多智能体系统的初步思考。

2 自底向上的TRUST分层评估框架

2.1 第1层：与任务一致的预测性、描述性和相关性（PDR）基础

> “连自己在说什么都不知道，精确又有什么意义？”——约翰·冯·诺伊曼（《物理科学中的方法》）

第一层也是最基础的一层，直接建立在 Murdoch 等人 [19] 提出的可解释性形式化定义之上。解释方法及其关联的人工智能系统至少必须满足三项要求，以确保其能够提供可靠且可在社会技术层面采取行动的见解：

[1]可使用包含 TRUST 卡的互动表格，网址为 https://forms.gle/jro7XjNhWvi6wDD18

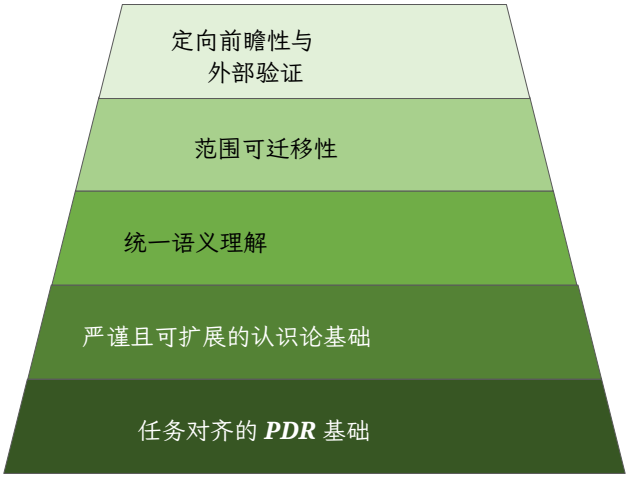


图1. 自下而上的 TRUST 评估框架

现实检验（*“P”*）。预测准确性的范围扩大，涵盖人工智能系统现实检验的力度，确保经过解释的系统能够捕捉数据中的真实内在关系。值得注意的是，准确性不应局限于标准的平均准确率。相反，评估需要采用多种准确性指标（例如敏感度、特异度、最坏情况误差、前 K 名），并针对与任务相适应的关键子群体进行分层评估。

描述准确性（*“D”*）。解释方法的输出必须忠实反映人工智能系统的内部决策逻辑。这意味着输出应能清晰地转述，并且易于目标利益相关者理解，同时不牺牲忠实度。

相关性（*“R”*）。这要求所传达的信息对于特定任务中的特定目标利益相关者确有用处。关键在于，任务和目标利益相关者的选择，与被解释的人工智能系统被赋予的目的，以及该系统将嵌入其中的人工工作流程直接相关。这对于可争辩性尤为重要：若描述性解释未纳入规范性基础，那么它对于争辩便毫无相关性，因为它无法赋能相关主体有意义地质疑或影响系统背后的价值对齐。

2.2 第2层：严谨且可扩展的认识论奠基

*“[*使用任何事物的人，凭借经验最了解它*...]*。例如，吹笛者会告诉制笛者，哪些笛子吹起来反应灵敏、适合演奏，并会要求制作所需的笛子；制笛者则会听从并为他效劳。”——柏拉图《理想国》第十卷

值得信赖的现任领域专家能够提供独特且至关重要的现实检验（PCS 中的“P”）。他们不仅拥有关于领域特定动态的独特情境知识，还负责确保人工智能系统在整合后不会破坏工作流程的目的 [18]。然而，随着数据可用性的增长，人工智能输出借助更多计算资源的扩展速度，远远超过了人类专家处理能力的提升速度。因此，详尽的逐实例审查变得不可行。这一扩展瓶颈直接关联到可计算性原则（PCS 中的“C”）；在此，该原则的范围从算法计算延伸至资源现实可行性，即人类时间

和认知努力方面。第2层“严谨且可扩展的认知论奠基”要求解释方法明确识别这些值得信赖的专家，并构建其输出，以便在大规模情境下开展系统且高效的审查。对于应对第2层而言，三项标准尤为重要。

客观专家验证。 用于审查解释方法输出的专家或其他参考方，必须是独立、权威且值得信赖的领域知识来源。在审查过程中，输出的呈现必须主动减轻确认偏误和自动化complacency。解释方法常常会通过为系统决策提供看似合理的解释而产生说服力风险。为应对这一问题，审查程序应强调受控方法，包括盲法专家评估以及阴性对照的呈现 [1, 9]。

认知兼容表征。 解释方法输出的呈现应努力与既有的领域术语或格式保持兼容（例如数学方程、逻辑陈述或标准化评分量表）。通过以熟悉且便于参考的格式呈现输出，有限的专家认知资源便可集中于关键决策逻辑，而不是解析新颖的呈现形式。然而，这种对兼容性的追求不应以阻止解释方法呈现挑战现有范式的新颖洞见形式为代价。

可扩展审计流程。 为专家审查选择数据或示例时，必须有意识地在专家现实可用的时间与严谨性之间取得平衡，并直接与系统的社会技术目的和优先事项保持一致（例如，接近模型决策边界的输出，或来自高风险子群体的输出）。此外，聚类技术可以将反复出现的逻辑归为一组，从而支持结构层面而非孤立实例层面的审查。在接受安全性和可信度测试后，专门的专家人工智能系统也可以集成进来，执行半自动化审计。

2.3 第3层：统一语义理解

“名不正，则言不顺。”——孔子（《论语》）

根据 PCS 的稳定性原则（“S”），第3层级——利益相关者间的统一语义理解——评估解释方法的输出能否为相关利益相关者提供明确且稳定的见解。

基线输出稳定性。从根本上说，要实现统一理解，解释方法的输出必须在任务情境中面对合理扰动时保持稳定且一致。使用同一内部模型处理相似输入时（即使数据存在合理扰动或随机噪声），也必须产生功能上相似的解释输出。否则，脆弱的解释方法必然会导致相互矛盾的结论。

语言稳定性。不同学科间的术语冲突尤其值得关注。对工程师而言，“矩阵”可能指表格数据，而对生物学家而言，它可能与细胞微环境相关。同样，对统计学家而言，“偏差”可能表示估计量的系统性偏离，而对监管者而言，则可能引发对公平性和歧视问题的担忧。解释方法必须使用各方共通且稳定的语言来传达其背后的决策过程，从而降低假设相互冲突的可能性。

代表性示例。 应在经过有意选择且清晰记录的代表性示例中，验证统一的语义理解。根据总体、关注问题、代表性与审查（PQRS）框架 [32],，代表性

用于评估所选示例与潜在部署用途中目标总体的接近程度。总体的统计学定义是：在现实世界中生成数据和输出的所有条件、环境及子群体的完整集合 [32]。值得注意的是，解释方法的使用方式可能与底层已部署人工智能系统的目标总体略有不同。解释往往会在后果重大的决策或边缘案例决策中被不成比例地调用，而在这些情形下，获得不同利益相关者的认同最为关键。应仔细记录正式的抽样策略，例如在领域关键子群体或决策边界区域中进行分层随机抽样，以便有力地证明所选示例能够代表目标总体（例如，记录抽样方法说明或代码、纳入／排除标准）。如果未采用随机抽样（基准中经常如此），则需提供文档，说明基准的目标总体，并阐明其如何实现代表性。

2.4 第4级：范围可迁移性

“预见 *consiste* 于尝试将我们在先前案例中观察到的内容应用于新案例；但新案例永远不会与先前案例完全相同。”——亨利·庞加莱（《科学与假设》）

稳定性原则（“S”）还要求，用于现实世界部署的解释方法能够稳健应对新的数据集来源以及相关领域，甚至相距甚远的领域。在部署时，人工智能系统将面临各种部署条件，因此也会面对不同的目标总体。数据整理人员可能采用不同的标注标准，或使用不同的硬件记录数据。工程师可能通过不同的流程管线存储和预处理这些数据。领域从业者可能采用不同的规程来规定人工智能系统的使用方式和时机。因此，了解解释方法在这些变化下是否仍然值得信赖至关重要。因此，第4级“范围可迁移性”评估的是：一种解释方法是否已证明能够超越其原始数据集和情境，泛化到更广泛的数据集类别，或相关甚至相距甚远的任务。

广泛验证。 从历史上看，监督学习模型通常需要针对至少几十个、甚至数百个基准数据集进行严格检验，以证明其有效性。相比之下，解释方法通常只在少数几个高度具体的示例上进行验证。第4级则要求在广泛的公共和私有数据集上进行实证验证，通过至少几十个多样化任务开展测试，以达到早期监督学习验证的规模。

适用范围界定。 没有任何单一的解释方法能够适用于所有可能的领域。尽管如此，在各种数据集上进行测试，可以为清晰界定该方法的适用领域提供证据。

受控环境验证。 第4级所涵盖的一种重要范围可迁移性形式，是在具有充分理解的真实值的受控条件下测试解释方法。在现实的合成环境中，或在经过简化但仍然现实的任务版本中，证明这些方法能够成功恢复真实值动态，可以为解释方法的性能提供证据，然后再据此证明其适用于更加开放的现实环境。

2.5 第5层：针对性前瞻性验证与外部验证

“经验不会出错　　. 只有你的判断会犯错，因为你期待她做到力所不能之事。”——列奥纳多·达·芬奇（《大西洋古抄本》）

第5层“定向前瞻性与外部验证”要求实现解释方法目前或许可以说是最终目标的事情：生成能够引导利益相关者发现新颖且可执行的知识，并通过定向前瞻性与外部验证加以证实的洞见。第4级侧重于回顾性数据集上的稳健性，已触及这一点；而第5层则更进一步，从观察性评估走向前瞻性真实世界执行。

可执行干预。 要通过第5层评估，解释方法的输出必须提出具体且可检验的目标，以供前瞻性干预。这可以包括新颖的定向数据收集、前瞻性湿实验室实验，或随机干预现场试验。

实证确认。 第5层还要求这些行动带来经实证确认的积极结果，其定义应与人工智能系统的根本目的及相关利益相关者密切相关。在科学发现领域，解释方法可能促成一项可重复率很高的实验，从而为一项新颖的科学定律或原理提供依据。在医疗保健或公共政策领域，解释方法可能促成改善健康结局或生活质量的干预。解释方法应通过清晰的记录，明确说明积极结果是什么，以及谁会从中受益。

部署认可。 从纯粹的算法洞见转向真实世界实验，隐含着对部署的强烈认可：提出并执行相关的物理实验，需要对运行动态和可行性有更深入的理解。通过第5层评估，意味着成功开展有意设计的外部前瞻性验证，这构成了最有力的证据，证明某种解释在目标驱动系统中的可执行相关性和真实世界稳定性。

3 生物学金标准：迭代随机森林（iRF）

TRUST框架广泛适用于各类人工智能系统。不过，其核心原则的灵感源自迭代随机森林（iRF） [5, 15, 28], 这一算法（及其变体）的开发；该算法通过提高随机森林的稳定性，支持发现可解释交互。评估iRF完整展示了TRUST从头到尾的流程。

iRF采用二分类设置，通过根据上一轮迭代的MDI重要性分数对分裂所用的特征进行加权抽样，为随机森林（RF）增加稳定性；第一轮迭代则使用RF。在森林中，同一路径上的两个特征被视为存在交互。随机交集树（RIT）遍历这些路径，并提取候选布尔特征交互。通过外层自助法为交互分配稳定性分数，并仅保留稳定的交互。

基础版iRF可识别单个特征的重要性及布尔特征交互。有符号iRF（siRF）纳入特征分裂的方向性，而低信号iRF（lo-siRF）则将流程调整为适用于低信号场景。我们在三种生物学场景中考察iRF：

果蝇胚胎发生 iRF 根据转录因子结合、组蛋白修饰和染色质可及性预测增强子状态，为发育生物学家提供指导 [5, 15]人类细胞中的可变剪接 iRF 根据组蛋白修饰和结合蛋白预测 RNA 剪接率，帮助生物学家揭示共转录调控机制[5]肥厚型心肌病（*HCM*） iRF 预测左心室质量（LVM），帮助心血管生物学家发现导致特定心脏病（HCM）的基因和多基因上位性相互作用 [28]。

3.1 第1层：任务对齐的 PDR 基础

iRF 在全部三种生物学场景中均满足 PDR基础。

"P" 在果蝇研究中，iRF 的表现与最先进的标准随机森林相当。在人类剪接研究中，iRF 将留出测试集上的性能从 0.44 提升至 0.51 AUC-PR，超过随机森林，同时将树的决策路径集中于更少的特征。在肥厚型心肌病研究中，由于疾病的复杂性，信号较弱；lo-siRF 在分类准确率上略微超过了其他基线方法，同时所使用的布尔交互更接近具有生物学依据的上位性相互作用 [28]。

"D" 外循环自助法机制能够识别稳定的布尔决策路径，直接反映预测引擎的内部机制。

"R" 在全部三项研究中，预测结果（增强子活性、剪接率、左心室质量表型）都是相关的生物学目标。同样，输入和解释方法的输出（表观遗传标志物、单核苷酸变异）也与每种结果所公认的生物学机制驱动因素相一致。决策树阈值化也更忠实地反映了生物分子相互作用过程中的生物学表达阈值。

3.2 第2层：严谨且可扩展的认识论基础

iRF 还通过专家有意开展的交互式审查，或与既有领域知识进行比较，来满足第2层要求。

严谨性 在果蝇研究中，iRF 揭示的 20 个二阶相互作用与既有文献进行了比较。在 HCM 研究中，揭示出的基因和基因-基因相互作用由心血管生物学家以盲法方式进行审查，并设置了阴性对照（排名居中的相互作用和随机相互作用）。

可扩展性 RIT+bootstrap 展示了算法可扩展性，成功应对了庞大的搜索空间（HCM 中有 $> 10^9$ 个候选项）。通过输出排名靠前且高度稳定、易于生物学家理解的布尔规则，为人工审查提供了便利。然而，规模极其庞大的候选项集合仍然对生物学家审查的真正可扩展性构成挑战。

3.3 第3层级：统一语义理解（利益相关者之间）

iRF 稳健地满足第3层级要求。iRF 和 RIT+bootstrap 都强调高度稳定的特征和输出，确保研究结果在不同运行中保持一致，从而具备可复现性。此外，使用具有领域意义的变量进行明确的布尔逻辑表达，可形成清晰的基于逻辑的研究发现，并在生物学家之间提供明确无歧义的含义。

3.4 第4级：范围可迁移性

iRF 也应对了范围可迁移性问题，但仍有必要将其扩展至更多数据集和领域。正如这三项研究所展示的，iRF 架构能够在不同生物学领域中实现泛化，从果蝇到人类心脏疾病均适用 [5, 28]。iRF 在简化的设置中也表现出良好的性能。在果蝇研究中，当增强子活性遵循局部尖峰稀疏（LSS）动态时，该框架表现出稳健性 [6]。与线性模型相比，lo-siRF 在纯加性遗传模拟中也能控制 I 型错误，并在上位性模拟（非线性基因-基因相互作用）中展现出更高的统计功效 [23, 28]。

3.5 第5层：有针对性的前瞻性和外部验证

iRF 最终稳健地解决了第5层问题。对于果蝇，20 个优先级最高的二阶相互作用中有 80% 已通过文献中的实验得到验证，而三阶和四阶相互作用则针对增强子活性提出了具体且可检验的假设 [5]。对于肥厚型心肌病，lo-siRF 提出的基因及基因交互中有 80%（5 个中有 4 个）随后通过物理湿实验

Tab表 1. TRUST 跨案例研究的评估摘要 研究

方法	第1层	第2层	第3层级	第4级	第5层
iRF	稳健	部分	稳健	部分	稳健
CD-T	部分	部分	部分	部分	未处理
SPEX	部分	部分	未处理	部分	未处理
GCT	部分	部分	部分	部分	稳健

实验 [23, 28]得到确认。

4 个其他解释方法案例研究

为进一步阐释TRUST框架，我们评估了其他解释方法。这些并非穷尽式评估，而是对其他解释范式的简要演示。

4.1 Transformer上下文分解（CD-T）

Transformer上下文分解（CD-T）是一种面向Transformer的内部机制解释方法，可捕捉输入特征组合或内部组件的贡献。在PCS的指导下，CD-T被用于分析一个经过病理报告微调的BERT模型，以对主要Gleason评分进行分类，这是一种癌症预后评分 [11]。具体而言，CD-T从一组病理报告中提取由注意力头组成的稳定电路。随后，CD-T通过识别该稳定电路的注意力图中经平均后获得最高关注度的词语，找出具有影响力的词语。其目的在于帮助泌尿科医生理解驱动Gleason评分分类的模型推理。主要利益相关者包括：审阅BERT预测结果并寻求临床依据的泌尿科医生、治疗路径取决于该癌症评分的患者，以及负责监督临床合规性的医疗管理人员。

从 TRUST层级的角度来看，CD-T 部分满足第1层至第3层的要求。对于第1层，底层BERT模型在一个已确立的前列腺癌评分分类任务上取得了较高准确率，从而支持 PDR 中的“R”和“P”。与路径补丁（第1层）相比，CD-T 还提供了更高的描述性忠实性（“D”）。CD-T 的输出经过泌尿科医生专家审核，结果显示，文本级归因突出显示了与 Gleason 相关的诊断标准（第2层），并帮助他们区分表现较好或较差的模型。CD-T 还以 PCS 为依据支持输出稳定性（第3层）。然而，一个语义鸿沟对这三个层级均产生了负面影响：注意力头电路体现的是计算流程，而非临床诊断依据，因此对非技术利益相关者而言含义模糊。在没有明确证据表明临床专家能够在这两个概念之间进行一致映射的情况下，专家审核的严谨性以及 CD-T 所带来的统一语义理解都受到质疑。

对于第4级，CD-T 具有很强的方法迁移性，能够跨 Transformer架构进行泛化，并在非临床文本（例如 SST-2、AGNews [11, 12]）上接受测试。然而，它无法基于这些输出促成或生成可执行的补救措施或具体的安全性防护，因此未达到第5层。

4.2 谱解释器（SPEX）

谱解释器（SPEX）专注于高效发现特征交互，尤其是在黑盒大语言模型中 [14]。SPEX 假设特征交互具有稀疏性，并利用稀疏图解码工具进行傅里叶分解，通过特征遮蔽高效找出主要的特征交互。为减少大语言模型调用次数，ProxySPEX [7] 通过使用代理梯度提升树（GBT）模型扩展 SPEX，利用特征交互的层次结构，实现

以十分之一的高昂大语言模型调用次数，达到相近的忠实性水平（强调 PCS 中的“C”）。此处我们探讨 SPEX 在 HotPotQA 数据集 [29],上的应用。该数据集要求综合不同文档中的事实，才能给出正确答案。此处的目的在于证明大语言模型确实整合并依据这些不同事实进行推理。利益相关者包括希望获得大语言模型正确推理证据的自然语言处理（NLP）和人工智能安全研究人员。

与 CD-T 类似，SPEX 在其构建和评估中部分覆盖了第1层至第2层，但持续存在的语义鸿沟削弱了其效果。对于第1层，经过评估的大语言模型在 HotPotQA 上取得了较高的基准精确匹配和 F1分数，从而确保了 PDR 中的“P”。对跨文档特征交互进行评估，直接针对所测试的多跳推理（“R”）。算法描述准确性（“D”）则通过实现更高的忠实性得到验证（比边际独立归因方法高出 20%）。然而，词语或短语之间复杂的布尔交互仍然难以供研究人员解释，也无法直接与现有知识源保持一致。事实上，对于第2层，认识论基础依赖于比较 SPEX 输出与人工标注的支持性事实之间的重叠部分。人工标注以词语平面集合的形式收集，这使人们质疑其与 SPEX 布尔交互输出之间的认知兼容性。这种歧义，加之缺乏交互式审查，进一步削弱了相关证据，即这些输出能够支持利益相关者之间统一的语义理解（第3层级）。

作为一种模型无关的黑盒解释方法，SPEX 从方法论角度来看具有很强的可迁移性，但目前仅在少数数据集上进行了测试。尽管 SPEX 提供了审计轨迹，有助于检测正确或错误的推理，但它未能进一步证明相关行动或可执行假设能够影响或改进模型推理，因此在第4层和第5层方面仍有所欠缺。

4.3 生成式因果测试

PCS 引导的生成式因果测试 [3] （GCT）是将第5层融入解释方法设计的典型范例。在神经科学领域，GCT 将复杂的脑编码模型转化为人类可读的假设，用于说明哪些特定概念会激活某些脑区。

GCT 训练一个预测性体素级编码模型，将大语言模型嵌入映射到功能性磁共振成像血氧水平依赖响应（用于捕捉 PDR 和 PCS 中的“P”）。随后，该模型在大规模语料库上运行，利用对比提示和自助法找出能够持续引发最强响应的 *n*-gram（PCS 中的“S”）。接着，GCT 输出一份由大语言模型引导的摘要，对这些 *n*-gram 进行概括，将其作为特定体素或脑区的候选假设。以这些输出为蓝图，大语言模型随后生成一个经过设计、能够触发目标体素的合成连贯故事，以确认或反驳所假设的概念。这里的主要利益相关者包括评估脑语言选择性的神经科学家。

从第1层开始，GCT 使用功能磁共振成像血氧水平依赖信号，符合认知神经科学的实践（“R”），并首先筛选出在留出测试数据上具有较高预测能力的体素（PDR 中的“P”）。然而，LLM 的中介作用损害了描述准确性，即 PDR 中的“D”。尽管基于语言的摘要更便于专家理解（支持第2层和第3层），但摘要所概括的概念可能受到大语言模型过度简化或产生幻觉的影响，从而切断其与底层模型逻辑之间的直接机制联系。在认识论基础（第2层）方面，GCT 根据先前发表的神经科学文献，独立找回了已确立的功能选择性（例如，纹外体区中的“身体部位”，或海马旁位置区中的“场景与环境”），但缺少交互式、有针对性的盲审流程。

在范围可迁移性（第4级）方面，GCT 被证明适用于不同尺度，从单个体素到更广泛的感兴趣皮层区域，包括在

前额叶皮层中发现的新微区域，但并未在种类繁多的数据集上运行。最值得注意的是，GCT 在设计中纳入了第5层，因此脱颖而出。GCT 明确设计了一项体内实验，以确认或证伪解释方法的输出。候选假设可作为蓝图，用于设计由大语言模型辅助生成的合成故事，以验证这些假设。由此产生的实际脑反应直接确认或证伪解释方法输出的因果有效性。这些体内实验的成功率各不相同，总体而言不及 iRF 的 80% 成功率。

5 讨论

我们提出 TRUST，将其作为一种有原则、分层次的人工智能评估框架。我们的核心愿景是，将人工智能解释方法提升为独立的评估指标，这对现代人工智能和数据知识发现至关重要。TRUST 在满足人工智能治理可争辩性这一基本原则的同时，牢牢立足于领域知识和可执行干预，从而支持以科学严谨、可信且可争辩的方式部署人工智能。

对解释方法进行严格评估，有助于实现人机协作。在我们看来，人工智能融入工作流程的理想方式，类似于数学与直觉之间由来已久的关系。人类凭借身体经验形成直觉，并据此指引创新方向。人工智能系统则提供计算引擎，用以验证、驳斥或完善这些直觉。人类直觉与数据驱动的人工智能验证并非各自为战；只要可信的可解释性发挥关键桥梁作用，两者之间的反复迭代就能开创大量新知。不过，提升可解释性并非要把人工智能的学习和发展限制在人类认知能力的范围内。数据驱动的人工智能模型有望发现人类无法企及的模式。

为此，尽管 TRUST 的实证基础以人机协作为核心，人工智能领域的发展促使我们设想将 TRUST 扩展至自主多智能体系统这一新兴范式。这种前瞻性调整需要有意识地扩展核心术语：利益相关者的范围应扩大，将机器协调者纳入其中；领域专家也应包括专门的评论智能体。相应地，评估一种解释方法将扩展为评估智能体间遥测，或评估多智能体网络中智能体之间的通信协议。值得注意的是，尽管可争辩性本质上以人为中心，但其透明性、质询和可引导性等核心机制，在这一智能体范式中同样至关重要，有助于机器协调者检测失调并防止静默故障级联。

无论解释方法是在人类直觉与人工智能之间建立桥梁，还是为自动化人工智能审计员提供支持，算法系统一旦部署于社会之中，就会成为社会技术系统，并需要具备可争辩性。人工智能系统旨在增强或取代现有实体，无论这些实体是成熟的流程、现任领域专家，还是承担责任的把关者。从实践角度来看，解释方法是现任者审查即将部署的人工智能系统的重要途径。分层的 TRUST 框架因此不可或缺：并非所有解释方法都同样值得信赖，能够满足部署就绪度要求；而自动化智能体系统则需要经过验证的遥测保真度。满足 TRUST 全部五个层级的解释技术，为安全部署、监控和质询人工智能系统奠定了必要基础。随着数据知识发现实践以及更广泛的人工智能部署格局不断演进，TRUST 提供了一套灵活的框架，可纳入更多层级，以维护扎根于领域的知识发现，以及负责任且可争辩的治理。

致谢

本研究得到了挪威研究委员会、TRUST—挪威可信人工智能中心项目（项目编号：357425）的支持。我们还衷心感谢以下机构提供的部分支持：

美国国家科学基金会（NSF）资助项目 DMS-2413265、NSF 资助项目 DMS 2209975、NSF 资助项目 DMS-2515767、NSF 关于合作研究：数据科学基础研究所（FODSI）的资助项目 2023505；美国国家科学基金会（NSF）和西蒙斯基金会通过项目编号 DMS-2031883 和 814639 对深度学习的理论基础合作研究提供的资助；以及美国国家科学基金会（NSF）向人工网络威胁情报与行动研究所（ACTION）提供的资助项目 MC2378。

参考文献

[1] Julius Adebayo、Justin Gilmer、Michael Muelly、Ian Goodfellow、Moritz Hardt 和 Been Kim。2018。显著性图的健全性检查。《神经信息处理系统进展》31（2018）。[2] All Things Nordic。2026。挪威为学校中的人工智能划定界限。访问于2026年9月。https://all-things-nordic.com/2026/06/22/norway-draws-a-line-on-ai-in-schools/ [3] Richard Antonello、Chandan Singh、Shailee Jain、Aliyah Hsu、Sihang Guo、Jianfeng Gao、Bin Yu 和 Alexander Huth。2024。生成式因果测试：在语言神经科学中连接数据驱动模型与科学理论。arXiv预印本 arXiv:2410.00812（2024）。[4] Kenneth A. Bamberger、Deirdre K. Mulligan 和 Victor Zhenyi Wang。2027。可争辩性：人工智能治理中认知上负责任的知识生产。《法律理论探究》28（2027）。（手稿）。[5] Sumanta Basu、Karl Kumbier、James B Brown 和 Bin Yu。2018。利用迭代随机森林发现预测性且稳定的高阶交互作用。《美国国家科学院院刊》115，8（2018），1943–1948。[6] Merle Behr、Yu Wang、Xiao Li 和 Bin Yu。2022。基于随机森林获得的树集成中可证明的布尔交互恢复。《美国国家科学院院刊》119，22（2022），e2118636119。[7] Landon Butler、Abhineet Agarwal、Justin Kang、Yigit Efe Erginbas、Bin Yu 和 Kannan Ramchandran。2026。ProxySPEX：通过大语言模型中的稀疏特征交互实现高效推理可解释性。《神经信息处理系统进展》38（2026），72306–72340。[8] Dara-Abasi Ita，《福布斯》。2026。那个指望无人上诉的算法。访问于2026年9月。https: //www.forbes.com/sites/daraabasiita/2026/06/09/the-algorithm-that-counted-on-no-one-appealing/[9] Susanne Gaube、Harini Suresh、Martina Raue、Alexander Merritt、Seth J Berkowitz、Eva Lermer、Joseph F Coughlin、John V Guttag、Errol Colak 和 Marzyeh Ghassemi。2021。人工智能说什么就做什么：临床决策辅助工具部署中的易感性。《NPJ数字医学》4，1（2021），31。[10] 州长加文·纽森。2026。州长纽森签署全美首项人工智能保障措施以保护加利福尼亚州居民，并呼吁联邦政府履行自身职责。访问于2026年9月。https://www.gov.ca.gov/2026/09/09/governor-newsom-signs-first-in-the-nation-ai-safeguards-to-protect-californians-calls-on-the-federal-government-to-do-its-part/ [11] Aliyah Hsu、Georgia Zhou、Yeshwanth Cherapanamjeri、Yaxuan Huang、Anobel Odisho、Peter Carroll 和 Bin Yu。2025。使用上下文分解在Transformer中实现高效的自动化电路发现。载于《国际学习表征会议》，第2025卷。57880–57898。[12] Aliyah R. Hsu、Yeshwanth Cherapanamjeri、Anobel Y. Odisho、Peter R. Carroll 和 Bin Yu。2024。通过Transformer中的上下文分解实现机制性解释。arXiv预印本 arXiv:2407.00886v1（2024）。arXiv:2407.00886v1 [cs.AI][13] 《商业内幕》。2026。ChatGPT扩大其在美国学校课堂中的影响力。访问于2026年9月。https: //www.businessinsider.com/openai-chatgpt-for-teachers-us-schools-2026-8[14] Justin Singh Kang、Landon Butler、Abhineet Agarwal、Yigit Efe Erginbas、Ramtin Pedarsani、Kannan Ramchandran 和 Bin Yu。2025。SPEX：扩展大语言模型的特征交互解释。arXiv预印本 arXiv:2502.13870（2025）。[15] Karl Kumbier、Sumanta Basu、Erwin Frise、Susan E Celniker、James B Brown 和 Bin Yu。2018。用于识别增强子相关转录因子结合的有符号迭代随机森林。arXiv预印本 arXiv:1810.07287（2018）。[16] Aaron J Li、Nicolas Sanchez、Hao Huang、Ruijiang Dong、Jaskaran Bains、Katrin Jaradeh、Zhen Xiang、Bo Li、Feng Liu、Aaron Kornblith 等。2026。绿色防护：迈向可信人工智能的以用户为中心的方法。arXiv预印本 arXiv:2604.24700（2026）。[17] Weixin Liang、Mert Yuksekgonul、Yining Mao、Eric Wu 和 James Zou。2023。GPT检测器对非英语母语写作者存在偏差。《Patterns》4，7（2023）。[18] Deirdre K Mulligan、Daniel Kluttz 和 Nitin Kohli。2019。塑造我们的工具：将可争辩性作为促进专业领域负责任算法决策的手段。SSRN 3311894（2019）。[19] W James Murdoch、Chandan Singh、Karl Kumbier、Reza Abbasi-Asl 和 Bin Yu。2019。可解释机器学习中的定义、方法与应用。《美国国家科学院院刊》116，44（2019），22071–22080。[20] NextGovFCW。2026。人工智能已经投入工作，美国人口普查局正在统计其应用情况。访问于2026年9月。https://www. nextgov.com/artificial-intelligence/2026/08/ai-already-job-and-census-keeping-count/415528/ [21] 英国国家医疗服务体系英格兰分部。2026。50万名英国国家医疗服务体系员工将获得新的人工智能工具，以帮助腾出更多时间服务患者。

访问于2026年9月。
https://www.england.nhs.uk/2026/06/500000-nhs-staff-to-get-new-artificial-intelligence-tools-to-help-free-up-more-time-for-patients/ [22] Zachary T Rewolinski 和 Bin Yu。2025。人工智能时代真实数据科学的PCS 工作流。*arXiv*预印本 *arXiv:2508.00835*（2025）。[23] Tiffany Tang。2023。用于生物医学问题可靠假设生成的统计机器学习。博士学位论文。加利福尼亚大学伯克利分校。[24] Philipp Tschandl、Christoph Rinner、Zoe Apalla、Giuseppe Argenziano、Noel Codella、Allan Halpern、Monika Janda、Aimilios Lallas、Caterina Longo、Josep Malvehy 等。2020。用于皮肤癌识别的人机协作。《自然·医学》26，8（2020），1229–1234。[25] 美国联邦地区法院，纽约南区。2023。Mata 诉 Avianca 公司案，678 F. Supp. 3d 443（S.D.N.Y. 2023）。[26] 美国人口普查局。2026。美国企业对人工智能的使用情况。访问于2026年9月。https://www.census.gov/library/stories/ 2026/05/ai-use-businesses.html[27] Kristen Vaccaro、Karrie Karahalios、Deirdre K Mulligan、Daniel Kluttz 和 Tad Hirsch。2019。算法系统中的可争辩性。载于*2019*年计算机支持的协同工作与社会计算会议配套出版物。523–527。[28] Qianru Wang、Tiffany M Tang、Michelle Youlton、Chad S Weldy、Ana M Kenney、Omer Ronen、J Weston Hughes、Elizabeth T Chin、Shirley C Sutton、Abhineet Agarwal 等。2025。上位性调控心脏肥大的遗传控制。《自然·心血管研究》4，6（2025），740–760。[29] Zhilin Yang、Peng Qi、Saizheng Zhang、Yoshua Bengio、William Cohen、Ruslan Salakhutdinov 和 Christopher D Manning。2018。HotpotQA：用于多样化、可解释多跳问答的数据集。载于*2018*年自然语言处理中的实证方法会议论文集。2369–2380。[30] Bin Yu。2013。稳定性。《伯努利》（2013），1484–1500。[31] Bin Yu 和 Rebecca L Barter。2024。真实数据科学：负责任的数据分析与决策实践。麻省理工学院出版社。[32] Bin Yu 和 Karl Kumbier。2018。人工智能与统计学。《信息技术与电子工程前沿》19，1（2018），6–9。[33] Bin Yu 和 Karl Kumbier。2020。真实数据科学。《美国国家科学院院刊》117，8（2020），3920–3929。