arXiv:2302.00799v1 [cs.HC] 1 Feb 2023

绘制可解释人工智能中的社会技术鸿沟：弥合可解释人工智能鸿沟的框架

UPOL EHSAN, 佐治亚理工学院，美国KOUSTUV SAHA,微软研究院，加拿大MUNMUN DE CHOUDHURY,佐治亚理工学院，美国
MARK O. RIEDL,佐治亚理工学院，美国

可解释人工智能（可解释人工智能）系统本质上具有社会技术属性，因此面临社会技术鸿沟——技术可供性与社会需求之间的差距。然而，绘制这一鸿沟颇具挑战。在可解释人工智能的情境下，我们认为，绘制鸿沟有助于加深对问题的理解，并能以此为基础反思，进而提供可操作的洞见，改善可解释性。我们利用两个不同领域的案例研究，通过实证推导出一个框架。该框架将可解释人工智能情境下的人工智能指南联系起来，并阐明如何运用这些指南弥合鸿沟，从而助力系统地绘制社会技术鸿沟。我们将该框架应用于一个新领域的第三个案例，展示其可供性。最后，我们讨论该框架的概念层面影响，分享其操作化过程中的实践考量，并就如何将其迁移至新的情境提供指导。该框架在增进对可解释人工智能中社会技术鸿沟的概念性和实践性理解方面作出了贡献，从而拓展了可解释人工智能设计空间。

ACM分类概念： • **以人为中心的计算 → 人机交互实证研究**；**用户研究**；协作与社会计算实证研究； • **计算方法 → 人工智能**。

附加关键词和短语：可解释人工智能、社会技术鸿沟、人机交互、框架、以人为中心的可解释人工智能、公平性、问责制、透明度与伦理、负责任人工智能、人工智能伦理、组织动态、人工智能治理、用户研究、参与式设计

ACM参考文献格式：
Upol Ehsan、Koustuv Saha、Munmun De Choudhury和Mark O. Riedl。2023年。绘制可解释人工智能中的社会技术鸿沟：解决可解释人工智能鸿沟的框架。*ACM*人机交互会议论文集7，CSCW1，第34篇文章（2023年4月），33 页。https://doi.org/10.1145/3579467

1 引言

可解释人工智能（XAI）系统在本质上具有社会技术属性，因为其技术（人工智能）组件嵌入在社会环境中 [3, 43, 87]。作为一个研究领域，可解释人工智能旨在为系统的行为提供人类能够理解的解释 [4, 48, 65]。鉴于其社会技术属性，且这类系统日益被部署于医疗保健 [29, 76, 78, 96], 、金融 [102, 117], 和刑事司法 [68, 80, 126], 等高风险领域，我们需要兼顾社会和技术因素，以减少偏见并促进可解释人工智能中的问责 [142]。

所有社会技术系统，无论是否属于可解释人工智能，都面临着Ackerman 所称的社会技术鸿沟——“我们知道在社会层面必须支持什么，与我们在技术层面能够支持什么之间的分野” [3]。Ackerman 将这一鸿沟视为社会技术系统面临的一项核心挑战，并强调若不正视这一鸿沟，我们就无法有效地设计出有用的系统 [3]。社会技术鸿沟最引人深思、也最具挑战性的一个方面，是难以彻底且永久地弥合它 [3, 41, 116]。这是因为“社会活动是流动且微妙的”（这意味着

作者地址：Upol Ehsan，佐治亚理工学院，美国佐治亚州亚特兰大，ehsanu@gatech.edu；Koustuv Saha，微软研究院，加拿大魁北克省蒙特利尔，koustuvsaha@microsoft.com；Munmun De Choudhury，佐治亚理工学院，美国佐治亚州亚特兰大，munmund@gatech.edu；Mark O. Riedl，佐治亚理工学院，美国佐治亚州亚特兰大，riedl@cc.gatech.edu。

2023年。2573-0142/2023年/4-第34篇
文章https://doi.org/10.1145/3579467

用户需求是动态的），而“技术系统僵化而脆弱”、“缺乏社会适应性”，并且“无法容纳足够的细微差异”，尤其是“在支持社会世界方面” [3]。

考虑到这些挑战，弥合差距乍看之下似乎徒劳无功：为什么要费心弥合一个无法填补的差距？然而，Ackerman提出了一个关键观点——提出这一差距的目的并非“解决”这个根本问题，而是深入理解它 [3, 41]。如果不理解它，我们就无法应对这一差距。如果不绘制或映射这一差距——也就是了解这一差距是什么样的，以及它的两侧分别是什么——我们就无法充分理解它并加以应对。以下类比或许有所帮助：想象两座山之间有一道峡谷。如果我们不了解这道峡谷所形成鸿沟的地形（例如，鸿沟最宽或最窄之处在哪里），就很难找到建造桥梁的最佳地点。我们需要一些机制来绘制或映射这一鸿沟。本文正是在可解释人工智能的情境下开展这项工作——提供一些机制，用于绘制可解释人工智能系统的技术维度和社会维度，以便我们更好地理解并应对其中的社会技术鸿沟。绘制这一鸿沟有助于更清晰地呈现设计空间，并促进明智的设计决策。

鉴于人工智能是一种“新颖且难以驾驭的设计材料” [42, 42, 61]，描绘可解释人工智能中的社会技术鸿沟颇具挑战，尤其是考虑到当前系统 [161, 162]具有非确定性、随机性和黑箱特性。这意味着，与非人工智能系统相比，描绘人工智能驱动系统的技术可供性时存在更大的不确定性。因此，在可解释人工智能系统的情境下（而非非人工智能系统），描绘这一鸿沟面临着转化方面的挑战。

受 Ackerman 目标的启发，本文提出，在应对可解释人工智能中的社会技术鸿沟时，应从“填补鸿沟”的理念转向“理解鸿沟”。我们主张，参与描绘鸿沟的过程能够加深我们对问题的理解，并反过来提供可操作的洞见，以提升可解释性。我们从概念和实践两个层面将这一主张付诸实施。在概念层面，我们对描绘可解释人工智能中社会技术鸿沟的过程进行定位。通过将重点转向理解问题的效用，这一过程有望揭示可解释人工智能领域中的认知盲点，进而拓展设计空间。在实践层面，它能够应对一个探索不足的领域：我们该如何着手映射这一鸿沟，并利用映射结果应对可解释性方面的挑战。我们利用来自不同领域的两个真实案例研究，在实证基础上推导出一个分析框架，以系统地描绘这一鸿沟，从而应对描绘可解释人工智能中社会技术鸿沟所面临的挑战。该框架不仅整合了人工智能和人机交互领域现有的指南，还将其应用拓展至可解释人工智能的情境，凸显社会与技术要素之间的相互作用。我们还探讨如何借助新兴的可解释人工智能概念，将框架的技术侧与社会侧联系起来；这些概念以社会技术视角审视可解释性。我们将所构建的框架应用于一个新领域的第三个案例研究，并展示两项主要成果：（1）更好地理解（描绘）鸿沟（即鸿沟的形态），从而能够识别鸿沟中最有希望和最具挑战性的区域；（2）提出应对鸿沟的可操作洞见（即如何利用鸿沟），进而反思性地提升用户可操作性方面的认知，并增强系统的可解释性。总而言之，我们的贡献有四点：

- 我们将社会技术鸿沟的绘制置于可解释人工智能的情境中，从而拓展设计空间
- 我们通过对不同领域中的两个真实案例研究进行实证分析，推导出一个用于绘制鸿沟的框架，并将其应用于一个新领域的第三个案例研究
- 我们的框架将人工智能和人机交互领域中多种多样的指南联系起来，并将其转化到可解释人工智能的情境中，阐明如何通过连接基础设施的社会层面与技术层面来利用这些指南弥合鸿沟
- 我们探讨该框架在概念层面的意义，分享其操作化过程中的实践考量，并就如何将其迁移到新的用例领域提供指导。

我们的工作旨在帮助利益相关者绘制可解释人工智能系统中社会技术鸿沟的轮廓。我们框架的设计哲学是生成式且迭代式的，而非规范性的 [33, 82, 107]。通过“从过程而非解决方案的角度重新聚焦设计” [137], ，它可以帮助利益相关者避免某些技术中心主义的陷阱，例如解决方案主义（总是寻求技术解决方案） [114]和形式主义（过度抽象并寻求数学解决方案） [63]。本文并非对可解释人工智能中社会技术鸿沟的详尽论述；相反，它通过将鸿沟绘制过程操作化，迈出了应对这一问题的基础性一步。

2 背景与相关工作

2.1 人机交互、计算机支持的协同工作与人工智能系统中的社会技术鸿沟

Walker 等人 将社会技术系统定义为包含“社会”与“技术”方面相互关联的系统。他们指出，“将‘社会’与‘技术’相结合的必然结果是，社会层面未必会像技术层面那样运作；人并非机器。矛盾的是，随着复杂性不断增加，就连‘技术’也可能开始表现出非线性行为 [153]。Ackerman在其广受赞誉的著作中 [3], 指出，计算机支持的协同工作与其技术机制的社会需求之间存在固有鸿沟；这一社会技术鸿沟关乎我们知道必须在社会层面支持什么，与我们在技术层面能够支持什么之间的分野。多年来，社会技术鸿沟的证据已在诸多人机交互和计算机支持的协同工作技术的不同方面与实例中显现，例如计算可供性和数据隐私 [17, 55, 141]。

基于人工智能的计算系统嵌入于社会情境中，研究也指出，技术中心主义视角会对人工智能造成不利影响 [127, 139, 149]。如果未能平衡地考虑社会因素和技术因素，人工智能系统就难以融入个人和组织的工作流程 [105, 129, 157]。此外，忽视社会因素可能导致潜在的误用、不信任 [25, 163, 164], 、伦理风险和意外后果 [111, 132, 147]。以往研究明确指出了在社会技术情境中构建基于人工智能的计算系统时涉及的抽象与陷阱 [10, 63, 114, 137]。值得注意的是，Selbst 等人探讨了技术设计者如何规避这些陷阱，并划定抽象边界，将社会行动者纳入其中，而非仅纳入技术因素 [137]。Andrus 等人探讨了为公共利益开发人工智能所涉及的诸多方面，以及社会技术如何与公平机器学习、人工智能安全和人在回路自主性这三个当代领域密切交织 [10]。

同样，Alter 发表了一篇概念性论文，旨在弥合将系统视为工具与将系统视为由人类参与者构成的社会技术系统之间的差距 [8]。 Munson 等人 借鉴了 Ackerman 在计算机支持的协同工作概念中提出的社会技术鸿沟，回顾了当代利用社交媒体促进健康的社会技术挑战 [116]。 Malatji 等人 考察了组织信息与网络安全实践中的社会技术鸿沟 [106]。

Ackerman 将社会技术鸿沟视为计算机支持的协同工作中的核心挑战 [3]。尽管人们已积极尝试应对可解释人工智能中的社会技术挑战， [38, 43, 44, 87, 154],但如何弥合这一鸿沟仍缺乏充分研究。据我们所知，这是首篇明确聚焦于绘制并弥合可解释人工智能系统情境中的社会技术鸿沟的论文。本文借鉴计算机支持的协同工作领域丰富的文献，案例研究聚焦于人工智能支持的工作，这类工作不仅具有合作 特征，也具有协作 层面（例如跨职能团队）。因此，为绘制可解释人工智能中的社会技术鸿沟，本文立足于计算机支持的协同工作社区长期以来设计和研究可解释计算系统的传统，并在此基础上加以拓展 [52, 69, 83, 93, 101, 122, 135]。

2.2 可解释人工智能及其社会技术层面

广义而言，如果人工智能系统旨在使其决策易于理解和解读为人们所理解 [12, 27, 48, 62, 66, 95, 108, 123]，那么就可以认为它具有可解释性（或属于可解释人工智能）。可解释性通常比模型透明度 [62, 95, 123]涵盖的范围更广。近期的可解释人工智能研究表明，事后解释 [48, 95] 和算法可审计性 [62] 能够建立人们对人工智能系统的信任。总的来说，可解释性是一种人为因素，而非仅仅是模型固有的属性 [12, 13, 45, 108, 112]。因此，近期研究倡导在可解释人工智能领域采用以用户为中心的方法 [47, 108, 139, 149]。Wang 等人回顾了决策理论，并指出可解释人工智能的输出在支持人类推理的完整认知过程方面存在诸多不足 [154]。

人机交互和计算机支持的协同工作领域长期以来一直关注如何使计算系统具备可解释性 [9, 50, 60, 79, 94, 100, 155]。Abdul 等人研究了人机交互领域关于可解释系统的研究，涵盖专家系统、推荐系统、情境感知技术和机器学习系统。Lim及其同事的一系列研究展示了人工智能可理解性在情境感知系统中的重要性、机遇与挑战 [90–92, 94]。Miller 指出，可解释人工智能应从以算法为中心转向以人为中心，并强调可解释人工智能算法输出与人们所寻求的解释特性之间存在差距 [108]。此外，Doshi-Velez和Kim指出，应在预期的使用情境中让目标利益相关者参与其中 [40]。

采用以人为中心的方法，近期研究探讨了可解释界面如何帮助模型开发者诊断和改进机器学习模型 [74, 79]。Kaur 等人考察了数据科学家如何频繁误用可解释性工具并对其过度信任 [79]。研究人员已在多个领域探索人工智能系统的可解释性功能，包括健康 [154, 159], 和招聘 [30]。多利益相关方参与的价值也已在内容审核 [84]和主动学习 [61]中得到关注。与此同时，研究人员已开始调查并评估可解释人工智能技术的有效性 [7, 23, 26, 39, 46, 160]。然而，可解释人工智能与人类解释的哲学与心理学基础之间存在脱节，这一点已得到充分指出 [110], , Miller呼吁借鉴社会科学的见解便是最好的例证 [108]。有趣的是，尽管可解释人工智能常被认为是迈向可问责人工智能的关键一步，但实证研究发现，几乎没有证据表明解释能够提升用户感知到的问责水平，或增强用户对人工智能系统的控制权 [121, 142]。鉴于可解释性涉及人的因素，且可解释人工智能系统具有社会技术属性，我们的工作拓展了可解释人工智能系统社会技术需求的设计空间。

2.3 将人工智能中的用户需求转化为可操作方案

随着人工智能系统日益融入社会，我们需要将最佳实践转化为可操作的措施并予以实施。基于此，Amershi等人提出了18条适用于人机交互的设计指南 [9]。关于可解释人工智能使用者的典型角色，已有若干分类法 [12,71, 148], , 旨在为应对人工智能算法的选择与转换所带来的挑战提供有价值的见解。此外，研究人员还探讨了以用户为中心的方法，以设计可解释人工智能用户体验 [51,87, 158]。Liao 等人研究了可解释人工智能相关的设计挑战，并提出了一种问题驱动的设计流程，以明确用户需求、可解释人工智能技术的选择、可解释人工智能用户体验的设计与评估 [87]。最近，Ehsan等人通过引入并探讨社会透明度这一概念，拓展了人工智能系统中可解释性的内涵。社会透明度是“一种以社会技术视角为基础，将社会组织背景纳入人工智能介导的决策解释之中的观点。” [43]作者考察了由四个W所承载的社会组织信息——谁做了什么、何时以及为何——如何通过设计特征增强人工智能的可解释性 [43]。为应对算法治理带来的社会问题，Lee等人提出了一个集体参与式框架，使人们能够为自己的社区制定算法政策 [86]。

既有研究也倡导或提出了使人工智能系统更以人为中心、并置于社会组织情境中的方法 [37]。Ehsan和Riedl采用批判性技术实践 [6] 和反思性设计 [138]，提出了反思性以人为中心的可解释人工智能（HCXAI）的基础。Madaio等人采用参与式方法，共同设计检查清单，以了解人工智能伦理方面的组织挑战 [103]。以及Zhu等人提出了价值敏感算法设计（VSD） [170]，通过在算法创建的早期阶段让利益相关者参与其中，避免设计选择中的偏见，或避免损害利益相关者的价值观 [170]。从“人工智能生命周期”的视角出发，Dhanorkar等人强调，不同类型解释受众的可解释性需求会随时间变化 [38]。既有研究还借助设计虚构和推测性情景，理解人工智能系统设计中的用户价值观和文化视角 [31, 32, 115]。Šabanović提出了相互塑造与共同生产框架 [127]。Jones等人提出了一种智能社会技术系统设计流程，同等重视对部署情境中社会概念的分析，以及以计算形式呈现这些概念。在技术可供性方面，Sanneman和Shah设计了一个三级框架，用于开发和评估可解释人工智能系统的行为 [133]。

在其他研究中，Gebru 等人提出了数据集数据表，旨在促进数据集创建者与数据集使用者之间更有效的沟通，并鼓励优先考虑透明度和问责 [59]。 Sokol 和 Flach提出了可解释性事实表，使研究人员和从业人员能够了解特定可解释人工智能方法的能力与局限 [143]。在模型方面，Mitchell 等人提出了模型卡，以鼓励围绕模型的创建、整理和评估进行透明的报告 [109]。在信任和人工智能伦理方面，Bellamy 等人、Hind 等人探索了如何在组织环境中将其付诸实践 [20, 72, 81]。

我们的工作在上述研究基础上进一步推进，并为其作出贡献，提出了一个框架，帮助识别、落实并弥合可解释人工智能系统及人工智能可解释性中各种社会和技术组成部分之间的差距。为此，我们不仅借鉴了上述观点，还将它们整合起来，以描绘可解释人工智能系统中的社会技术鸿沟，并在基础设施层面弥合这一鸿沟。

3 可解释人工智能在现实世界中的应用案例研究

我们分享了来自不同领域的两个案例研究：销售（第3.1节）和心理健康（第3.2节）。这些案例研究展示了真实世界中的项目，这些项目由一位或多位作者参与，历时两年。对于每个案例研究，背景部分都会提供必要的情境、问题、社会与技术组成部分之间相互作用的描述（这些内容有助于界定差距），以及我们为解决已描绘出的社会技术鸿沟所采取的步骤。在概述这些相互作用时，我们聚焦于相关方面，最终据此构建一个框架（第4节），帮助我们梳理新型可解释人工智能使用场景中的社会技术鸿沟。详尽罗列技术属性和社会需求既无益处，也无必要，甚至可能根本无法做到（该列表可能大到不可数）。

与其自上而下地先介绍框架，不如采用自下而上的“逐步构建”方式，先构建与情境相关的部分，再在之后将其与相关概念联系起来，这或许能让读者更容易理解。这种方式也反映了我们如何在现实生活中逐步形成这一框架。我们通过第一个案例研究，为框架构成模块所依据的数据和动机奠定基础，并以此为视角阐述第二个案例。这样的结构或许有助于读者将这一过程迁移到其他案例中。我们以一种能够作为绘制社会技术鸿沟之构成模块的方式，来组织这些要素之间的相互作用，并最终据此开发框架。我们不仅分享了两个案例研究，以经验性地推导出框架，还通过另一个不同领域（网络安全）的真实案例研究应用了该框架（第5节）。我们提供了一个端到端的、

从构想至实施的叙述，说明绘制鸿沟如何有助于理解并应对它。我们还讨论了如何将该框架迁移到其他领域，详见第6节

3.1 案例研究1：Kuro，人工智能驱动的销售软件

3.1.1 背景。SalesCorp（化名）是一家《财富》美国500强跨国科技公司，投资了500万美元开发人工智能定价推荐系统，以协助其科技销售团队。每笔销售的金额通常在50万美元至200万美元之间，因此风险很高。如果客户的预算无法负担报价，SalesCorp可能会失去眼前的这笔交易，并面临失去该客户长期收入的风险。SalesCorp采用企业对企业（B2B）模式，销售硬件、软件和云基础设施。它采用动态定价模型，该模型取决于多个因素，例如客户的历史预算、公司规模、合作关系的持续时间等。

人工智能系统Kuro（化名）被介绍为“一位随时待命的虚拟资深同事”，并声称利用超过50,000个数据点来生成推荐价格。该系统通过七三合作开发：70%由内部团队利用专有数据开发，30%则与提供人工智能基础设施支持的第三方供应商合作开发（例如，获准独家使用的专有模型）。人工智能驱动的Kuro所承诺的价值主张之一是其鲁棒性，据称相较于其“脆弱”的前代系统Sumo有所提升。

3.1.2 问题。表面上看，Kuro 的指标相当出色：准确性达到 90%，模型透明度良好（展示排名前 3 的特征），并在每条建议旁显示置信度范围（0–100%）。然而，实际情况却不尽如人意——只有 10% 的销售人员使用 Kuro；而在使用者中，只有 2% 认为它有用。鉴于部署所处的社会组织背景，这里的问题并非纯粹的技术问题；它具有很强的社会层面因素（下文将对此展开说明）。考虑到 Kuro 面临的问题，SalesCorp 聘请了本文的一位作者来处理这一问题。

在为期一年的合作过程中，我们将在下文说明并阐述相关方法和数据，这些最终构成了框架的社会与技术两翼的基石。请注意，我们对以下事件的叙述结构进行了安排，以便读者能够在下文的案例研究与下一节提出的框架之间找到对应关系。尽管我们在文中依次介绍这两个方面，但它们实际上大体上是并行推进的，并通过反复迭代、相互促进，逐步清晰地呈现出社会技术鸿沟。为保护客户保密性，引文中的专有细节均已删去。

3.1.3 为框架构建模块奠定数据、流程与动机基础。 认识到基础设施的社会技术属性后，我们聚焦于社会技术鸿沟的两个相关方面：理解技术可供性和社会组织层面的需求。为了解这两个方面，我们采取了以下方法。

采用参与式方法 [136], ，我们举办了两场研讨会，每场有 28 位参与者。参与者来自不同岗位（5 位可解释人工智能工程师、6 位产品经理、5 位负责任人工智能经理、3 位数据科学家、4 位技术副总裁和 5 位用户体验研究员）。我们采用参与式方法，以利用利益相关者的领域专长，并确保他们的声音在干预措施中得到积极体现。在说明会上，我们向利益相关者介绍了社会技术鸿沟这一概念，以便聚焦于当前系统在技术上支持什么与我们在社会层面必须支持什么之间的差异。他们中近半数此前已了解社会技术鸿沟这一概念。此外，通过介绍相关工作的示例 [38, 43, 88, 89], ，我们引导他们从以人为中心的可解释人工智能视角出发；这种视角不会将可解释人工智能局限于模型透明度，而是将社会技术因素纳入其概念化过程 [45]。这有助于参与者在解决问题时采取社会技术立场（而非技术中心主义立场）。

下面我们分享各场研讨会的**关键要点**，以此为数据提供依据，并说明框架的结构安排。

研讨会*1*：构建构成模块。这场为期3小时的研讨会旨在集思广益，并就思考问题的技术和社会方面时应考虑的相关构成模块达成共识（这些模块后来构成了最终框架的基础，详见第4节）。

研讨会的前半部分是一项创意生成（头脑风暴）活动，四个跨职能团队通过便利贴列出相关主题（构成模块）。在这一阶段，我们明确要求参与者专注于提出想法，而非评估这些想法。各团队在技术侧提出了 7 个区块，在社会侧提出了 12 个。

后半部分，我们着重以兼顾实际可行性与概念严谨性的方式精简构成模块。四个团队都认为，19 个区块总数太多，难以在实践中发挥作用。接下来，我们将团队两两配对——四个团队合并为两个；这两个团队分别达成共识后，再合并成一个大团队。由于每个团队都由跨职能成员组成，讨论以积极的方式挑战了既有假设。一位数据科学家这样说道：“让自己的观点受到质疑，感觉相当耳目一新。我以前从未真正想过自己为什么需要了解社会层面的内容。这个过程让我清楚地看到，自己为什么会陷入视野狭窄的状态。”通过反复迭代的亲和图法，我们为技术侧归纳出三个主题区块，为社会侧也归纳出三个（我们并未要求两侧的区块数量必须平衡）。

最后一项任务是为这些区块确定名称。我们请参与者分享每个区块的主要功能。在**技术**方面，他们列出了：**数据**谱系（数据的来源、收集方式等）、**模型**/算法的功能可供性（模型能够或不能够做什么），以及人工智能系统生成**解释**的方式。在**社会**方面，他们概括了：**信任**（用户在多大程度上依赖人工智能）、解释的**可操作性**（用户需要什么，才能根据人工智能生成的解释采取行动），以及组织层面的**价值观**（组织激励与价值观如何和个人期望相互作用）。最终，小组确定了这些区块：技术方面包括数据、模型和解释；社会方面包括信任、可操作性和价值观。我们将这些构成模块用作分析视角，以组织案例研究，同时也将其作为框架的基础（第4节）。

研讨会*2*：各构成模块包含什么。本次为期4小时的研讨会旨在系统地将每个区块转化为可实际运用的内容。我们可以使用哪些问题或指南，实际探究每个区块，从而加深对社会技术鸿沟的理解？第一步是协作探讨一系列问题或指南，以充实每个区块背后的理念。为实现这一目标，我们最初为每个区块设计了定制问卷。然而，在没有充分利用现有工作流程和流程的情况下引入新材料，遭到了强烈反对。例如，公司已有相应流程，由工程师和数据科学团队准备过程数据文档，例如《数据集数据表》 [59]。一位产品经理提到采纳新问卷可能遇到的阻力，并尖锐地问道：“还没探索我们能在多大程度上利用现有资源之前，谁会同意增加新的工作？”根据这一反馈，我们转而利用现有流程和文档。与第一次研讨会类似，我们将参与者分成四个跨职能团队，共同负责找出与不同构成模块相关的现有流程和指南。接着，我们将团队两两配对，以便反复筛选，直至达成共识（四个团队合并为两个，最后再合并为一个）。各团队得以指出重要内容。例如，团队一致认为，模型模块可以使用《模型卡》 [109]，数据模块可以使用《数据集数据表》。

第二步是确定每个区块下应包含多少个问题或指南，以及如何将它们归类。为了兼顾实际可行性并减轻负担，我们收到的反馈是不要生成详尽的问题清单。正如一位数据科学家所说：“给我们一些入门问题[，来自不同的指南]，以此开启对话。.. [并且] 让我们能够从中筛选问题

剩余的内容”（着重号为后加）。对于每个区块，我们采用主题分析 [22], 与参与者合作，纳入不同指南中的相关“入门”问题。接下来，我们运用亲和图法 [19, 21] 与利益相关者合作，将每个主题下各个分支的相关问题归类（例如，数据集数据表中的哪些问题最相关）。最后，我们征求了 6 位可解释人工智能和人机交互专家的反馈，以完善最终的入门问题清单。

这种奠基构成了最终框架的基础（第 4 节）。在此理解的基础上，我们现在分享如何探索技术和社会组织可供性，以解决 Kuro 面临的问题。

3.1.4 理解技术可供性。要剖析技术分支，需考察三个构成模块：数据、模型和可解释性。

对于**数据**，我们采用参与式方法，借鉴《数据集数据表》中的问卷 [59]，并针对公司进行了定制，以收集有关数据谱系的见解。

对于**模型**，由于供应商提供了部分决策机制，我们遇到了一些“禁区”。在可能的情况下，我们访谈了供应商的客户经理，并根据我们的使用场景调整了《模型卡》 [109]。虽然无法完成原始模板中的每一项，但空白项和已填写数据的项目同样有用。它明确提醒我们有哪些信息尚缺，最终帮助我们界定技术可供性的范围。

对于**可解释性**，我们主要依靠两种制品。首先，我们从Liao 等人的可解释人工智能问题库中选取了相关问题 [87]。结合数据表和模型卡，这些问题为我们探索不同方面提供了良好的起点，例如解释类型（如何、为什么、假设情境、为何不等）。其次，我们改编了 IBM 的可解释性事实表 [20]，以了解功能和验证层面上的解释生成机制（模态、全局与局部等）。由于原始事实表包含一份由 36 项组成的工作清单，我们只能使用其中一部分。正如一位开发人员所说：“要做的事太多，得到的价值却太少。”

3.1.5 理解社会组织需求。社会层面的工作与技术层面的工作同步开展。我们希望明确了解参与度为何只有 10%。为剖析这一问题，我们结合访谈、愿景卡练习 [57], 和研讨会，着重了解三方面：对人工智能的信任（哪些因素会损害信任？哪些因素有助于建立信任？还需要什么？）、解释的可操作性（哪些因素可以帮助用户根据人工智能的决策采取知情行动），以及价值观（与组织规范之间的张力和一致之处）。在调查**信任**时，我们发现，尽管模型透明度有所帮助，但不足以校准对人工智能的信任。经验较少的销售人员往往觉得这款价值数百万美元的 Kuro“无所不能”，因此过度信任人工智能，并觉得自己没有发言权或自主权来反对人工智能。他们没有让自己陷入这种处境，而是选择完全忽视人工智能，不与之互动。经验更丰富的销售人员则只是忽略人工智能的建议，因为他们觉得自己“更懂”。理解信任的细微差别有助于我们更好地界定社会需求。

考虑到**可操作性**，Kuro提供的解释信息被认为是“惰性的”。一位销售人员说：“我不知道该如何处理这些数字，[Kuro] 给我看的这些数字说明了什么？它的置信度是如何得出的？置信度是相对于什么而言的？”此外，假定的交互范式也有误——它设定的是1名人类对1个人工智能，而销售活动是群体性的（多对多）。许多销售人员通过模拟手段绕开了这种孤岛式约束。他们会走到团队中的其他成员面前（每位客户通常由一个由5名不同层级的销售人员组成的团队服务），进行咨询。

在考虑**价值观**时，我们发现组织规范中存在多种张力。推翻 Kuro 的决定极其繁琐——相关审计和升级处理给销售人员造成了过重的负担，因此他们一直对此置之不理。此外，初级销售人员并不清楚对 Kuro 的价格进行哪些调整才算“可接受”。销售人员认为

技术人员是“入侵者”，是想通过自动化让销售人员失业的人。这造成了紧张的关系，阻碍了利益相关者之间的合作。

3.1.6 应对差距。 与项目刚启动时相比，这些流程更清晰地呈现了社会技术鸿沟的两面，也凸显了利用设计搭建桥梁的机会。尽管我们承认差距将始终存在，但目标从来不是“填补”它，而是理解其边界，并通过Ackerman 所说的一阶近似来应对它——“能够解决特定问题一部分、且权衡已知的可行方案” [3]。

将信任、可操作性和价值冲突联系起来的一个核心启示是，销售人员希望了解其他人如何依据人工智能的建议采取行动。他们希望能以“外围视野”了解以往的用户轨迹。例如，初级销售人员会想知道资深同事如何使用人工智能系统，以及为什么这么做。销售人员认为，这些额外的情境信息有助于他们更好地判断人工智能，让解释变得可操作。

我们测试了加入社会组织背景是否确实能有所帮助。我们临时采用了一种零散的方式：请销售人员将一份 Google 表格（在线协作电子表格）固定在 Slack 频道中，并填写以下信息：对于 Kuro 给出的任何处于临界状态的建议，他们都要记录事件 ID、自己采取的行动（接受/拒绝），并写下一段推文长度的理由说明（销售人员很喜欢把“为什么”浓缩成“推文”）。一周内，Google 表格中就有了 400 多条记录。销售人员甚至沿用了原有格式，为有用的评论添加“+1”（点赞）投票。我们发现参与度从 10% 上升到了 42%。这尤其令人惊讶，因为销售人员需要跳出原有工作流程，才能参与 Google 表格中的互动。访谈显示，他们认为这种“外围视野”足够有价值，值得克服上下文切换负担。

有了概念验证，我们便可以提出工程改进方案。通过参与式设计活动，我们设计了一项干预措施：除了 Kuro 的人工智能推荐和技术透明度之外，销售人员还可以看到其他人做了什么、何时做的，以及为什么。在开发这项干预措施时，我们遵循社会透明度框架 [43]，以促进对人工智能的信任校准，并确保集体行动的一致性。

实施后，我们仔细监测了参与度，发现其已提高至 87%，从 10% 大幅跃升。此外，销售人员表示，他们现在能更好地理解 Kuro，因为人工智能的决策如今置于人类行动的情境之中。新增的情境信息营造出一种整体性的可解释性，而这单靠技术透明度是无法实现的。通过后续访谈、问卷调查和研讨会，我们进一步完善了对差距边界的理解，使此前未被意识到的方面进入意识层面，从而变得可以采取行动。我们的行动也以反思性的方式使 Kuro 的解释变得可操作，提升了其可解释性。

3.2 案例研究2：人工智能驱动的临床心理健康工具

3.2.1 背景。 精神疾病的全球负担占伤残生存年数的32%，使这些疾病成为全球疾病负担的主要贡献因素 [152]。虽然及时治疗十分有益，但心理健康治疗在很大程度上依赖于患者在面对面咨询期间向临床医生所告知的内容。然而，临床医生经常报告称，患者的自我报告与外部现实之间存在矛盾 [15]。Fisher和Appelbaum [54] 提供了病例报告，其中临床医生开始将患者的电子通信（如社交媒体）纳入考量，作为一种新形式的旁证信息。社会计算和人机交互领域的既有研究表明，基于此类数据，人工智能和机器学习能够帮助推断某个人是否易受心理健康问题影响 [28, 67, 130, 131], , 例如重度抑郁症 [35] 和产后抑郁症 [36]。数字心理健康领域新兴的大量研究还论述了，如何利用人工智能系统所 harness 的计算能力来揭示复杂的

精神障碍的精神病理学，从而更好地为治疗应用以及不同临床医生之间的合作提供信息 [85, 165]。

3.2.2 问题。尽管许多患者已经会在就诊时向临床医生提供社交媒体数据，而患者 [16], 目前却没有技术支持这些互动，尤其是因为临床团队的不同成员对同类数据有不同需求。为了支持在诊疗现场发挥基于社交媒体的人工智能洞见的潜力，本案例研究中，美国东北部一家大型、城市化且享有全国声誉的医疗系统的临床研究人员和从业人员，与本文的一位作者 [165, 166]开展了合作。他们的核心目标是探究：将人工智能应用于患者自愿提供的这些数据，能否提供不同维度的视图（例如通过仪表板），以满足不同治疗团队成员各异的需求。

该工具的原型通过为期一年的合作，采用协同设计方法开发，并利用已同意参与的心理健康患者自愿分享的社交媒体（Facebook）数据；随后，研究人员在这家大型医疗系统中，通过一对一访谈（N=13）和焦点小组（N=13），邀请不同类型的心理健康临床医生对其进行了测试。本案例研究展现了该原型的潜力，同时也凸显了诸如临床医生使用这类“黑箱”工具所面临的责任问题。他们设想了这样的情形：尽管能够获取显示症状加剧的患者数据，临床医生却可能无法采取行动。与前一个案例研究一样，我们概述了该社会和技术基础设施的组成部分。

3.2.3 理解技术可供性。访谈和焦点小组揭示了多种技术特征，临床医生认为，从支持为患者提供协作治疗的角度来看，这些特征十分重要。延续前一个案例研究所采用的方式，我们围绕数据、模型和可解释性这三个维度，讨论这些技术可供性。

聚焦于**数据**，该研究强调了心理健康症状和临床风险因素相关的重要问题。临床医生参与者认为该原型很有前景；然而，他们希望该工具能够促进共同识别和可视化患者心理健康状况的各种不同指标，因为当患者与精神科医生、治疗师和社会工作者互动时，这些指标对不同临床医生的价值可能有所不同。

对于**模型**，参与者指出，该原型提供的人工智能衍生心理健康洞察存在可信度问题。有两个主要的可信度问题可能会削弱信任：社交媒体数据是否与现实生活相关（构念效度），以及在这些数据之上应用的人工智能是否能够区分特定帖子背后的不同情境和意图（临床效用）。

最后，谈到**可解释性**时，参与者表示，该工具有助于比较和讨论社交媒体数据所显示的内容与患者口头报告的内容，尤其是在两者相互矛盾时。它还能帮助一位临床医生了解团队另一位成员如何以及为何采用某些基于社交媒体的人工智能洞见，同时选择忽略其他洞见。要满足这些技术需求，就必须在工具中加入可解释性功能。

3.2.4 理解社会组织需求。与技术需求相辅相成的是，访谈和焦点小组还发现，要使基于人工智能的工具在诊疗现场发挥作用，还需要满足一系列社会和组织层面的需求。我们将在下文讨论这些需求。

临床医生参与者对使用该工具时的**信任** 提出了诸多顾虑。如果无法对算法的运行建立某种默认程度的信任，临床医生参与者认为，他们会谨慎看待基于社交媒体的人工智能洞察。值得注意的是，参与者担心，人们可能会因观察者效应 [34, 128], 而对该工具的能力和临床实用性失去信任：患者知道临床医生会关注并获取这些信息后，可能会停止在社交媒体上发帖，或开始自我审查。

谈到**可操作性**，不同的临床医生也可能以不同方式运用该工具呈现的信息：治疗师可以将这些数据作为家庭成员报告的补充，因为人工智能洞察所谓的“客观性”可能有助于厘清并整合不同的辅助信息来源，从而更准确地了解患者的状况。此外，一些参与者推测，该工具或许能发挥一定的预防作用：如果可以将患者住院治疗前的某种症状模式（例如自杀意念）与当前模式进行比较，这或许有助于开展预防性的危机干预服务。

最后，研究发现了若干价值冲突，无论是患者与临床医生之间，还是临床团队不同成员之间。关于临床医生彼此之间的价值张力，案例研究揭示了负担方面的潜在问题。参与者担心使用该工具可能增加工作时长和工作要求带来的负担，其中某些类型的临床医生可能比其他人承受更多负担。其他人则认为，在咨询前查看患者的社交媒体信息未必总是可行，因为患者数量可能非常庞大。负担过重的临床医生（例如社会工作者）通常采用不同于为少数患者提供药物治疗支持的临床医生的工作方式和价值体系。

3.2.5 弥合差距。现在，我们概述一些在本案例研究情境中关注社会技术鸿沟的可能干预措施。与第一个案例研究一样，这些干预措施旨在理解技术需求与社会需求的边界 [3], 而非试图将其“填补”。

首先，许多临床团队会定期安排同步会议，开展治疗方案的同行评审，并讨论标准治疗中的差异。如果我们整合来自患者社交媒体数据的人工智能驱动洞察，就需要一种方法来对其进行总结、综合和可视化，以供日后使用。为促进合作，我们可以设想未来的迭代版本将这些人工智能驱动的洞察与会议记录一并整合到电子健康记录（EHR）中，以供日后查阅。对工具设计进行这些调整，有助于让讨论更透明，也能提升人工智能的可操作性。

其次，我们发现，原型中缺少关于不同临床医生如何依据人工智能洞察采取行动的具体流程。例如，如果某个趋势显示症状有所加重，而治疗师发现精神科医生并未使用这项信息，那么这种信息差距可能会妨碍治疗师制定治疗计划、造成困惑，并削弱对工具的信任。因此，临床医生参与者强烈认为，未来需要能够支持这些社交互动的透明度。例如，我们可以设计经过筛选的视图，展示与特定患者相关的决策路径，并根据患者的具体症状或风险因素进行呈现。

4 描绘可解释人工智能中的社会技术鸿沟的框架

在此，我们介绍我们在可解释人工智能领域的设计视角、框架的构建、推导过程及其构成模块。

4.1 我们在可解释人工智能领域的设计视角：以人为本的可解释人工智能

鉴于我们的目标是描绘可解释人工智能中的社会技术鸿沟，我们采用一种融入社会技术视角的可解释人工智能设计理念，即以人为中心的可解释人工智能（HCXAI） [38, 45, 89]。HCXAI 将可解释性的概念拓展至算法边界之外 [43]，并将其视为一种取决于关系和受众的构念，而非模型固有的属性 [12, 13, 108, 112]。鉴于人工智能系统（人机组合体 [38, 43]）存在于社会技术环境中 [88, 146], ，仅有算法透明度不足以使其具备可解释性 [49, 108]。因此，要解释“黑箱内部”正在发生什么，往往还需要了解“黑箱外部”的情况 [38, 87],，这就要求我们考量整个人工智能生命周期（而不只是算法）。例如，人脸识别系统为何会对有色人种女性产生不成比例的错误分类 [24] ，可以通过考察训练数据中的人口构成来解释。以人为中心的可解释人工智能领域的新兴研究 [69, 120, 135]

展示了更广泛的可解释人工智能视角如何有望回应针对流行的以算法为中心的可解释人工智能技术的批评；这些技术可能不起作用 [7, 119, 169] ，并且可能带来风险 [79, 145]。

4.2 推导框架的总体设置

我们在第3节中分享了数据基础，以及构建模块背后的流程与动机。该框架首先通过两个侧面明确承认社会技术鸿沟：技术侧和**社会侧**。图1 展示了这一构想——每个侧面都包含三个相互关联的构建模块。技术侧包括数据、模型和人工智能生成的解释。社会侧的构建模块则是对人工智能的信任、解释的可操作性和价值观。为了落实每个模块的目标，我们列出了每个步骤中可能涉及的利益相关者（表1），并提供了包含方法学建议和现有指南见解的“入门资料包”。图1 中每个模块旁边提供的入门问题不应被视为面面俱到，因为现实世界的人工智能部署不存在放之四海而皆准的公式——流程如何展开取决于情境、领域和目标受众。这些示例问题旨在启发思路，并为绘制鸿沟的生成过程开个头。每个模块的问题都是结合我们的案例研究，根据现有指南（如上文所示）协作确定范围的。正如 Ackerman 所指出的，用户需求（社会侧）具有动态性，这正是社会技术鸿沟难以完全弥合的主要原因 [3, 41, 116]。因此，问题往往源于社会侧；例如，因人工智能的建议与价值观发生冲突，用户便难以据此采取行动。

下面我们将介绍各个构建模块的操作化方式，先从技术侧开始，再介绍社会侧。正如我们在案例研究中（第 3 节）强调的那样，尽管我们在行文中依次介绍这两个侧面，但它们彼此相连，并通过迭代相互构建，从而更清晰地呈现社会技术鸿沟。各模块在图示或行文中的排列顺序，并不影响其操作化顺序（我们可以来回切换、转换侧面等）。模块的遍历顺序取决于问题以及各阶段的发现；例如，框架应用案例研究（第 5 节）将展示这一过程如何从可操作性模块（社会侧）开始，再转向数据模块（技术侧）。介绍完各个构建模块后，我们将说明如何借助主题标签（标签）在模块之间移动，以便跟踪迭代过程。接下来，为了弥合差距，我们将分享以缓解措施 和 一阶近似 [3]形式提出的策略。最后，我们将介绍参与这一过程可能产生哪些最终产出。除了框架在映射和绘制方面的作用外，其核心贡献还在于将现有的人工智能和可解释人工智能指南融入社会技术鸿沟这一情境之中。

4.3 技术侧

框架的这一侧旨在界定系统能够承担和无法承担的范围。我们将这一空间划分为三个相互关联的区块：**数据、模型**以及（由人工智能生成的）**解释**。这一划分有两个动因：首先，从理论上看，这一划分位于现有指南所采用的可解释人工智能系统技术基础设施范畴内（例如，模型卡 [109], 人工智能公平性 360 [20], XAI 问题库 [87], 等）。其次，这一划分以我们的案例研究为实证依据；研究发现，它在分析上的明确区分与实践中的操作化之间取得了平衡，并得到利益相关者的认同。这种契合促进了相关指南的可迁移性，从而有助于各构建模块的操作化（详见下文）。

数据：该构建模块的目标是了解数据谱系、可供性和范围，以及数据能够或不能告诉我们什么（在可获得或可能的最大程度上）。数据基础设施是

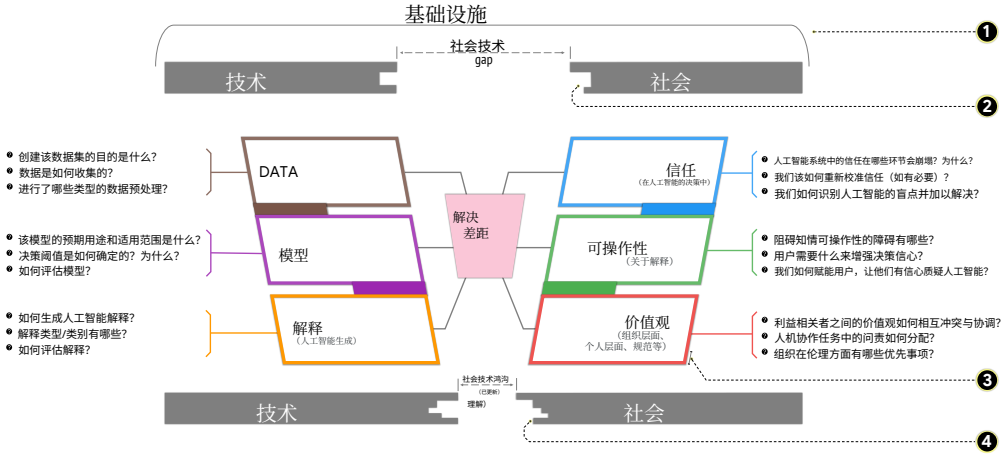

图1. 展示框架及其对社会技术鸿沟的影响（在可解释人工智能情境下）之前和之后这一过程。**(1)** 在顶部，我们从由社会侧和技术侧（灰色方框）共同构成的基础设施开始，社会技术鸿沟位于两者之间。**(2)** 在绘制鸿沟之前，我们对其边界形态只有一个“初步”的概念，图中以锯齿边缘表示。如何描绘其边界。在中间，我们通过将构成模块操作化，对鸿沟进行绘制，并“深入观察”每一侧。模块的顺序并不表示或限定任何先后次序。正如我们在第5节应用框架时将看到的，遍历各模块的顺序取决于具体问题以及我们从每个模块中获得的洞见。**(3)** 每一侧的倾斜意在表明，通过参与制图过程，我们能够获得可操作的洞见来弥合鸿沟（在图示上表现为两侧彼此靠近）。**(4)** 最后，在底部，经过制图过程后，我们对鸿沟的理解将更加“高分辨率”，图中以社会侧和技术侧更精细的锯齿边缘表示（不同于顶部的灰色方框）。映射后的视图揭示了两侧距离较近（凸显了值得着手弥合鸿沟的潜在领域）和距离较远（较具挑战性的领域）之处。总体而言，顶部表示映射过程之前对鸿沟的初步认识；中间表示过程中对两侧分别进行“深入观察”；底部则展示经过细化并充分绘制的鸿沟。每个模块旁还提供了示例“入门问题”，以启动其操作化过程。

模型的食物，这意味着它对整个生命周期具有根本性影响。因此，我们需要一份可靠且立足于具体情境的数据营养标签。为实现这一目标，我们可以借鉴并采用现有指南，例如数据集数据表 [59] 和人工智能事实表 360 [11], 以及数据集营养标签 [75]。利用现有指南，并在需要时进行调整，我们总体上关注数据集的起源故事。图1 展示了一些可用于了解创建目的、数据收集持续时间、预处理类型、数据集的新近程度、抽样信息等的入门问题。这些入门问题往往会启发更具情境针对性的后续问题。在获取数据谱系时，平衡广度与深度至关重要。该过程的利益相关者因组织而异，但通常包括数据科学家、数据分析师、数据库工程师等。

模型：本区块旨在界定模型的可供性与局限性，这对可信度至关重要。模型是将数据转化为有价值内容（如预测、分类等）的实体。为将目标付诸实践，我们可以借鉴模型卡等指南 [109]。图1围绕模型的预期用途、架构、性能评估以及用于评估的数据集，提供了一些入门问题。这些入门问题旨在引发深入的生成式

表1. 可解释人工智能社会技术鸿沟框架概述。

构建模块	Goal	操作化示例-ples	利益相关者示例
技术侧			
Data	了解数据谱系、可供性，以及数据的范围，了解数据能告诉我们什么。	数据集数据表 [59]，AI FactSheets 360 [11]，数据集营养标签 [75]	数据科学家，数据分-析师、数据库工程师
模型	界定模型的能力范围和局限性，这对可信度至关重要。	模型卡 [109]	Data 科学家，AI 运营释 工程师，产品经理
解释（人工智能生成）	界定.....的适用范围与局限人工智能生成的解释，尤其是围绕生成技术，模态和类别gory.	XAI 问题库 [87]	数据科学家，可解释人工智能工程师-程师、最终用户
社会侧			
（对人工智能的）信任	为了了解当前的信任以及校准信任的范围。	问卷调查 [14, 73, 140]，访谈由启动问题引导的问题，以及价值敏感设计（VSD）练习 [58]	最终用户，包括销售人人员和客户经理
可操作性（解释）	了解用户需要什么，才能以知情的方式根据人工智能的解释采取行动。以知情的方式。	社会透明度框架——架构 [43]	系统的最终用户（例如，分析师或心理健康专业人士）
价值观	为了理解组织层面的价值及其与个体期望之间的相互作用个人预期。	价值敏感设计练习、参与式设计工作坊	多利益相关者 涉及——其中包括 领导层、HR，开发人员、and end-用户·

讨论，聚焦于特定使用场景。利益相关者各不相同，但通常包括数据科学家、人工智能运维工程师（负责验证和测试模型的人员）、产品经理等。

解释：在此，我们希望界定人工智能生成的解释的能力与局限，尤其是生成技术、模态和类别方面。我们采用了广义定义，即解释是对“为什么”这一问题的回答 [97, 98, 108, 156]。为了 将这一目标付诸实践，我们可以调整现有的一些指南，使其适用于当前的使用场景。例如，我们可以借鉴 Liao 等人的 XAI 问题库中的部分内容 [87]，该问题库提供了解释类别（全局、局部、反事实等）的分类体系，列出了各自的生成技术，并将其与不同的问题类型（如何、为什么、如果.....会怎样、为什么不等）相联系，以促进用户研究。图 1 展示了部分入门问题，可用于界定人工智能生成的解释这一领域。 利益相关者通常包括数据科学家、可解释人工智能工程师（负责设计解释）和最终用户（依据解释采取行动）。

4.4 社会维度

框架的这一部分旨在梳理利益相关者的社会需求。我们将这一部分划分为三个相互关联的区块：**对人工智能的信任**（针对技术）、**可操作性**（个人层面）以及**价值观**（组织层面）。作此划分有两个原因：首先，它以我们的案例研究为实证依据；其次，它也与人工智能中的社会透明性所揭示的三个情境层面（技术、个人、组织）相一致。人工智能中的社会透明性这一视角将社会组织背景纳入人工智能可解释性的构想之中 [43]。这种一致性十分重要，因为社会透明度将有助于连接这两个部分（稍后详述），并解决社会技术鸿沟。

正如Ackerman 所指出的，技术可供性（系统能够做什么）往往是静态的，而社会需求则是动态的。用户需求的动态性正是社会技术鸿沟难以完全弥合的主要原因。问题往往出现在社会层面——例如，人工智能系统可能受到过度信任（技术层面），用户可能难以依据人工智能解释采取行动（个人层面），利益相关者之间也可能存在价值冲突等。

（对人工智能的）信任：此处的目标是了解当前对人工智能的信任水平，并确定我们如何恰当地校准信任。我们希望了解特定情境下用户对人工智能信任的基线理解，并找出信任崩溃的节点，以便采取行动。为将这一目标付诸实践，我们可以利用现有的对人工智能的信任问卷调查，例如 [14, 73, 140],由入门问题引导的访谈，以及价值敏感设计（VSD）[56] 练习，以了解人工智能性能的哪些方面对用户很重要、哪些方面缺乏信任，以及不同因素如何影响信任校准（何时信任或不信任人工智能的决策）等。图1展示了一些可用于引导调查的入门问题。此流程的利益相关者主要是需要根据人工智能的决策采取行动的最终用户。在 Kuro 的案例中，他们是销售人员或客户经理。虽然与最终用户合作很重要，但与数据科学家和开发人员（次要利益相关者）协作也很重要，因为他们通常负责进行更改。

（解释的）可操作性：此区块旨在了解用户需要什么，才能在充分知情的情况下依据人工智能的解释采取行动。可操作性指用户如何根据人工智能生成的解释采取行动，它在决策层面发挥着作用 [43]。有时，即使可解释人工智能系统运行良好，用户仍会觉得自己无力果断采取行动。可操作性处于对技术的信任与组织价值观的交汇处，彼此冲突的价值体系可能会降低人工智能解释的可操作性。回想 Kuro 案例，为了推翻人工智能的决定而进行的审计增加了负担。为了落实这些目标，我们可以借鉴社会透明度框架所提供的见解。该框架阐明了社会认同（来自同侪）、拓展对过往决策的“外围视野”，以及对后续行动的支持，如何提升可操作性和人工智能可争议性（图1 展示了一些入门问题）。为了了解这些构想应如何体现在系统设计中，我们可以采用与信任区块类似的方法工具——访谈和价值敏感设计练习。利益相关者通常包括系统的最终用户（例如，人工智能驱动的网络安全系统中的分析师，或我们使用场景中的心理健康专家）。

价值观：此处的目标是理解组织价值观及其与个人期望之间的相互作用。回想 Kuro 案例，组织看重的内容与销售人员的诉求之间存在错位，最终导致基础设施崩溃。为了将这一目标付诸实践，我们可以运用价值敏感设计（VSD）练习、参与式设计工作坊等方法工具，梳理组织与个人之间的冲突点和价值观一致之处，了解人机协作中的问责分配、工作绩效预期，以及对人工智能伦理的立场等（图1 列出了一些入门问题）。这是一个多利益相关者参与的过程，参与者包括但不限于管理者（领导层）、人力资源专业人员、最终用户和开发人员。

在绘制差距的过程中，应当考察所有构成模块，但并未规定考察各模块的先后顺序。每个模块都会受到其他模块的影响，不仅是同一侧的模块，也包括差距另一侧的模块。正如前文所述，考察顺序取决于具体问题，并由洞见驱动。解决可行动性模块时获得的洞见，可能会促使我们转而考察数据模块，从而更新我们的理解，并促使我们再次回到可行动性模块。我们可以在各个模块之间反复迭代，不断更新理解，直至对理解达到稳定饱和 [134] ，也就是进一步探究不再产生新洞见。为跟踪进展与成果，我们提供三个主题标签——基线理解、更新后的

理解和设计建议。基线理解（BU）指首次考察某个模块时，依据框架中的指南对其进行分析，以了解现状。更新后的理解（UU）指基于此前过程（例如，对另一个模块的基线理解）获得的新洞见，再次考察该模块时形成的理解。设计建议（DR）为工具、技术和政策的实际开发提供指导（例如一阶近似）。设计建议既可以源自基线理解，也可以源自更新后的理解。

4.5 弥合差距：缓解措施与一阶近似

请记住，没有灵丹妙药能够彻底弥合社会技术鸿沟 [116]。此外，弥合差距并不意味着做到尽善尽美。为应对这一差距，我们可以采用两种方法，即 Ackerman 所称的缓解措施和一阶近似 [3]。这两种方法并不相互排斥，可以结合使用。

缓解措施是指在政治、意识形态和教育层面采取的干预措施，这些措施会影响人们对系统可供性 [3, 116]的认知。其目的在于缓和差距，而不对系统进行系统层面的改变。这些措施可以包括利益相关者分析，例如我们在案例研究中开展的价值敏感设计实践。研究发现可以揭示系统中的盲点，进而帮助人们在使用系统时做出明智的决定 [116]。正如我们在 Kuro 的案例研究中所见，教育举措 [99, 104]）可以在技术力有未逮之处帮助弥合差距。正如案例研究所表明的，仅有一个称职的系统并不足以让人们认为它有用。结合分析，了解不同组织利益相关者在技术采纳生命周期 [18, 113], 中所处的位置，我们可以运用技术接受模型 [150, 151] 来理解人们对系统有用性的感知。这些步骤可以缓和社会与技术要素之间的某些脱节，并为未来的技术干预（如一阶近似）提供参考。

一阶近似是可处理的解决方案，能够部分解决特定问题，但存在已知的权衡 [3, 116]。要使一阶解决方案取得成效，可以将社会层面与技术层面连接起来，从而有效地应对一些最紧迫的社会技术问题，同时面对已知的权衡。正如案例研究所示，开发一阶近似有多种策略。在 Kuro 的案例中，我们展示了如何将社会透明度（ST） [43]作为一阶近似，以提供亟需的“外围视野”并促进参与度。通过呈现其他人做了什么、何时做、为何而做（四个W），社会透明度在三个层面（技术、个人和组织）呈现情境，从而促进可解释性；这三个层面与社会与技术两翼的三个区块相对应。社会透明度植根于可解释人工智能的理念，因此可以作为一种有效的一阶近似，用于弥合可解释人工智能系统的差距。借助社会透明度，提高社会技术因素的可见度能够带来可操作的洞见，进而有望弥合这一差距。为了将社会透明度付诸实践，我们可以利用其构成性设计要素——四个W（谁做了什么、何时做、为何而做）——来涵盖相关的社会组织背景。

4.6 框架成果

我们的框架有两个主要成果或最终产出：(1) 加深对差距的理解（差距是什么样的），以及 (2) 提供解决差距的可操作的洞见（如何应对差距）。上述主题标签与这些成果相互关联 —— BU（基线理解）和 UU（更新后的理解）对应第一个成果，而 DR（设计建议）对应第二个成果。我们的框架为复杂社会技术景观的可行探索提供指导，并产生解决差距的可操作的洞见。我们不仅应当了解最有希望搭建桥梁的地方，也应当了解差距最大的地方。因此，利益相关者能够就哪些差距需要应对、哪些不需要应对作出知情决策——这种能力正是 Ackerman 呼吁解决差距的核心要旨之一 [3]。

5 应用该框架

现在，我们分享第三个案例研究，其中将这一框架应用于现实世界的实际场景。为了展示该框架的多功能性，我们为应用选择了一个新的领域：网络安全。以下内容涵盖了为期 1.5 年的实地研究。在介绍应用细节之后，我们还将分享一项研究，用于考察该框架的有效性（利益相关者是否认为它有用？）。

5.1 背景：Atlas，一种人工智能驱动的网络安全系统

SecureCorp（化名）是一家《财富》美国500强跨国网络安全科技公司，投资了1000万美元打造Atlas（化名），这是一种人工智能驱动的网络安全系统，旨在帮助分析师管理防火墙配置，尤其是人们忘记关闭开放端口时出现的“臃肿”问题。随着组织不断发展，人们会忘记关闭开放端口，导致配置臃肿并造成严重的安全漏洞。这关系重大——未关闭的端口可能导致安全 breach，使公司蒙受巨额损失（数百万美元）；SecureCorp的客户主要是银行。纸面上看，Atlas表现出色，据称准确性令人称道（92%）；然而，用户参与度仅为2%。考虑到对Atlas的巨额投资及其内部部署，SecureCorp原本期待更高的参与度。由于参与使用Atlas的员工不到2%，问题是：为什么?

5.2 问题识别与使用框架进行对齐

我们的框架不仅帮助我们解决问题，还能为系统地着手 解决问题提供支持。识别问题后，我们需要将问题与框架对齐：判断它属于哪个侧面（技术侧或社会侧），并找到最匹配的构建模块。我们跟踪观察了10位网络安全分析师（下文简称分析师）两周，识别出问题——用户不太愿意 根据人工智能的解释采取行动。对照框架后，我们发现该问题属于社会侧中的**可操作性** 区块。

在接下来的 Atlas 案例研究中探讨各个概念区块时，我们使用基线理解 （BU）、更新后的理解 （UU）和设计建议 （DR）这些标签来跟踪洞见的进展。

5.3 基线理解（BU）

如第 4.6 节所述，基线理解提供有关问题的形成性信息。它通常出现在首次探访某个区块时，此时的探访就像是一次实地调查。

可操作性（社会层面）。为了获得基线理解，我们以框架中的入门问题为引导，对分析师进行了访谈。我们发现，分析师不愿根据这些解释采取行动，因为他们感到自己很脆弱，而且被孤立在真空中，无法与其他同事交流。工作本身包含合作环节，而 Atlas 无法支持这些环节。Julie（化名）的故事广为人知：她因听从人工智能错误地建议关闭某些端口（假阳性）而遭到解雇。然而，人工智能并未因此承担任何后果。

信任（社会侧）。这种脆弱性源于缺乏**对人工智能的信任**（位于图1社会维度的顶部）。在该框架的指导下，我们开展了一系列价值敏感设计（VSD）练习 [56, 58]，以了解基线，掌握他们对信任看重的方面。我们发现了两点，为后续步骤提供了依据：第一，问责在人机关系中被不公平地转移。讨论 Julie 的故事时，另一位分析师 Simon（化名）分享了关键顾虑：“人工智能不受影响……它没有任何损失。如果人工智能出错，而我信任它，那遭殃的就是我。所以，我为什么要信任一个不承担任何风险的东西？”第二，分析师们

他们无法信任人工智能，因为他们不了解技术基础设施。正如 Simon 所说："[他们] 真的不知道正在发生什么。"

这些见解促使我们转向技术侧，以便深入了解其组成部分的基线理解。我们使用了针对每个区块提供的指南（详见第 4 节），从实际操作层面界定技术可供性的范围及其局限性。

数据（技术）。针对**数据**区块，我们依据围绕数据来源、收集目的、政策等内容构建的框架，采访了系统管理员和数据科学家。当开发团队了解到分析师缺乏信任和可操作性时，他们决定参照《数据集数据表》框架 [59]开展定制化练习。我们获取了可供采取行动的信息；例如，该数据集包含专有信息，最初是出于合规原因收集的，而且关于预处理步骤的文档非常少，这促使开发团队着手弥补这一不足。

模型（技术）。对于**模型**，数据科学团队指出了一个重大局限——该模型属于专有产品，购自供应商（第三方）。因此，SecureCorp 的团队无法"打开"这个"黑箱"。我们还了解到，这种情况相当普遍——大多数组织都缺乏端到端构建系统所需的内部基础设施。企业通常会组合多种技术来提供服务。然而，制图过程揭示了盲点，这些盲点严重影响了信任和可操作性。当他们意识到自己无法回答我们的入门问题时，开发团队便有了明确的解决步骤——原本未知的问题变得清晰可见。

解释 （技术）。接下来，我们探讨了来自 Atlas 的**解释**。我们依据入门问题，与可解释人工智能工程师和分析师（最终用户）开展了多场焦点小组讨论。我们了解到，Atlas 会生成事后解释（即在模型训练完成后生成解释） [48, 95, 108, 124, 167] ，这些解释是局部的（解释某个预测，而非解释整个模型）[64, 125, 168]。这些解释经过了针对公司专有数据的微调。解释生成机制与模型无关，以适应无法访问供应商模型的情况。

尽管从社会侧和技术模块获得的这些见解颇有参考价值，但如果它们未能与组织价值观相协调，我们就无法据此采取行动；这也让我们回到了社会侧。

价值观 （社会）。为了对组织价值观与个人价值观之间的相互作用建立基线理解，我们依据框架指南，通过与领导层、人力资源代表、产品经理、分析师和审计员开展多场访谈和焦点小组讨论，找出一致与冲突之处。我们发现问责方面存在价值观冲突；例如，组织的设置使责任落在了人的肩上，这影响了可操作性（核心问题）。在此过程中，我们成功运用了前几个步骤中的见解，增进了同理心与理解，从而缓解了不同价值体系之间的冲突。

5.4 更新后的理解（UU）

基线理解促使我们重新审视某些区块，以更新我们的理解，而这一过程往往会产生设计建议。为清晰起见，我们在本小节中分享更新后的理解，并在下一小节中集中报告设计建议。我们先从重新审视数据 和模型开始。

数据（技术）。与审计员的交流（在<样式 id='3'>价值观的业务单元期间）产生了两项应用于数据模块的重要洞见。首先，我们了解到，与此前的理解相反，该数据集可以每年更新（而不是每 3 年更新一次）。其次，审计员指出，缺乏预处理文档会带来合规风险（这一问题在数据业务单元期间被强调）

我们通过焦点小组讨论将这些发现传达给开发团队，最终形成了具体的设计建议（将在下一小节中分享）。

模型（技术侧）。领导层在业务单元进行价值观讨论时得知，分析师无法了解模型的内部运作方式，于是与供应商重新协商了责任条款，从而减轻了分析师的负担，更新了我们对模型模块的理解，并提出了设计建议（详见后文）。

借助这些更新后的见解，我们重新审视了信任和可操作性区块，以及社会侧中的相关内容。

信任（社会侧）。请回想一下，了解技术侧业务单元的主要原因，是由于不信任人工智能，而这种不信任源于对技术模块缺乏了解。我们向分析师传达了关于数据、模型和解释的业务单元信息及向下理解。让他们更清楚地了解哪些事情能做、哪些不能做，本身就重新校准了他们对人工智能的**信任**。例如，了解新的数据更新政策后，他们对人工智能决策在应对最新安全威胁时的正确性形成了更准确的预期。与供应商重新协商模型责任，则减轻了他们的心理负担。

可操作性（社会侧）。鉴于信任与**可操作性**密切相关，信任的重新校准也更新了可操作性。请回想一下，分析师曾感到脆弱且彼此隔绝。我们希望解决这一问题。在框架的指引下，我们运用社会透明度（ST）的视角，通过提供过往决策的情境来拓展决策的“外围视野”。访谈表明，分析师非常认同这样一种做法：通过提高彼此互动的可见度来分担问责，从而降低个人的脆弱感。我们依据社会透明度的指南开展参与式设计会议，将社会透明度中四个W（谁在何时因何做了什么）如何在设计层面体现并促进知情可操作性这一问题付诸实践。遵循社会透明度的指南，我们了解到，增强社会组织背景能够促进同伴学习，并为个人决策增添社会认同。这些工作促成了设计建议（详见下一小节）。

经过这些讨论后，我们对问题的理解已达到饱和，[134]进一步调查也未能产生新的信息。随后，我们着手提出设计建议。

5.5 设计建议（DR）

更　p新的 由更新后的理解形成的设计建议涵盖技术层面和政策层面 .

在技术层面，建议涉及**解释**和**数据**。在**解释**方面，分析师希望将四个W设计洞见（源自可行动性的UU）与人工智能的解释一并纳入。对于每条人工智能生成的解释，他们都希望附上社会组织背景——过去还有哪些人基于类似的人工智能决策采取了什么行动，以及原因是什么（四个W）。四个W有具体的设计呈现方式：关于谁，分析师希望展示工作年限，以说明其建议背后的权威性。他们投票反对披露姓名，以保护隐私。关于什么，他们希望不仅显示过去的决策结果（接受/拒绝），还显示最终结果（威胁是否已被消除）。关于何时，除日期外，他们还希望提供更多与时间相关的信息，例如数据集最近更新时间。关于为什么，他们共同商定采用标准化模板，以确保一致性。在**数据**方面，鉴于对更新政策和预处理细节有了新的认识，这些信息应与其他详细信息一并标注，并归入四个W中的何时。

在政策方面，在**价值**业务单元期间，我们与领导层进行了富有成效的合作。我们的发现赋予了他们采取行动的力量，他们发起了一项由分析师（最终用户）主导的教育活动（一项姑息措施 [3, 116]），并据此改革了政策和培训流程。在**模型** 方面，我们修订了员工培训计划，以涵盖与供应商重新协商的责任合同，从而

确保新分析师不会遭遇 Julie 的命运。这些政策变更符合 Ackerman 提出的方法之一——教育干预，以应对社会技术鸿沟 [3]。

5.6 框架成果

为了协调用户需求、组织激励和技术可供性，我们与开发团队合作，实施了设计建议（社会透明度的四个W以及政策变更）。部署一个月内，参与度达到*93%*，与最初的*2%*相比有了大幅提升。在接下来的四个月里，我们通过访谈、问卷调查和研讨会持续追踪进展。结果令人鼓舞——参与度有所提升（平均为94%），解决了可操作性问题。与Kuro案例研究类似，社会透明度（ST）的四个W（谁做了什么、何时做的以及为什么）所提供的增强型“外围视野” [43] 帮助分析师更深入地理解社会组织背景，从而恰当地校准他们对人工智能的信任以及采取行动时的自信程度。技术可供性的可见度提高，使他们能够对Atlas的局限性形成知情判断。最重要的是，分析师表示，Atlas的可解释性有所提升，因为它不再缺少知情可操作性所必需的社会组织背景。

5.7 框架有效性的调查

尽管提升参与度对 SecureCorp 很重要，但作为一个研究团队，我们想要探究以下问题：该框架的有效性如何？利益相关者是否认为该框架有用？为回答这些问题，我们按照以下流程开展了一项后续研究。

5.7.1 方法。 我们举办了一场为期两天的研讨会，共有 23 位背景各异的参与者出席（下文称为 P1-P23），包括可解释人工智能工程师、产品经理、用户体验研究人员、人力资源管理者、负责任人工智能经理等。研讨会每天持续 5 小时。我们从主题分析的角度分析了 10 小时的录制影像 [22]。我们采用开放式编码方案，生成了原位编码（直接源自数据），并将其归类为主题。我们反复迭代，直到编码及相应主题趋于稳定，并结合思维导图和亲和图法，按主题层级对其进行分组。

5.7.2 研究发现。 下文将介绍我们关于框架有效性的研究发现、相关挑战，以及我们可能如何应对这些挑战。

该框架帮助参与者认识到问题具有多利益相关者和社会技术属性，从而拓展了他们对可解释性的理解。参与者强调，该框架帮助他们认识到，“一切都处于特定情境之中，没有什么是孤立发生的，我们都需要相互依赖”（P3）。对于新近形成的社会技术视角，具有工程背景的参与者表示，他们很庆幸意识到“这是一场多人游戏，而 [他们] 却把它当成单人游戏来玩”（P2）。他们也欣慰地表示，自己“不必独自应对人工智能的所有复杂问题，例如公平性、偏见和可解释性”（P23）。“该框架帮助 [他们] 认识到，要求工程师承担人工智能可解释性的全部重担是不公平的，因为这个问题不只是技术问题，更是社会技术问题”（P14）。最重要的是，社会技术视角让参与者能够将“可解释人工智能视为超越模型透明度的事物”（P8）。这位参与者道出了这一新发现：

> “很遗憾，我曾对可解释人工智能抱有如此狭隘的看法，认为可解释人工智能就等同于算法透明度。人工智能系统由人和算法构成。我现在意识到，系统的可解释性 远不止模型的可解释性……这就是为什么像我这样的 [工程师] 需要了解数据来自哪里、模型有哪些做不到的事……否则，我怎么能帮助最终用户知道何时该相信人工智能的解释，何时不该相信呢？”（P12，可解释人工智能工程师，重点为后加）

表2。 框架有效性研讨会参与者 详细信息

ID	组织角色	经验
	工程与数据科学	
P1	可解释人工智能工程师	>5 年。
P7	可解释人工智能工程师	>10 年。
P12	可解释人工智能工程师	>4 年
P23	可解释人工智能工程师	>3 年
P2	数据科学家	>10 年
P6	数据科学家	>5 ys.
P14	数据科学家	>6 ys.
	用户研究	
P3	用户体验研究员	>5 年
P4	用户体验研究员	>5 年
P15	用户体验研究员	>10 年
P16	用户体验设计师	>13 ys.
	产品管理	
P5	产品经理	>15 ys.
P8	产品经理	>8 年
P9	产品经理	>11 年
	组织治理	
P10	负责任人工智能经理/ 人工智能伦理	>3 年。
P11	负责任人工智能经理/人工智能伦理	>4 年。
P19	负责任人工智能经理/人工智能伦理	>3 年。
P20	负责任人工智能经理/人工智能伦理	>5 年
P13	合规官	>10 年
P22	合规官	>11 年
P17	技术副总裁	>15 年。
P21	技术高级副总裁	>20 年。
P18	人力资源经理	>16 年。

参与者对可解释人工智能中的社会技术鸿沟有了更深入的理解。在参与者接触这一框架之前，他们“并不清楚这一鸿沟是什么，也不了解技术和社会方面为何重要”（P10）。明确聚焦社会技术层面，并“为每一方面设置专门区块，使这一过程模块化且可追溯”（P14）。当被问及该框架如何加深他们对这一鸿沟的理解时，这位参与者做了很好的总结：

> “这个过程让我们清晰地看到了差距，而不是像刚开始时那样，只看到一幅模糊不清的图景。我们知道差距最大和最小的地方在哪里。这些知识对于制定干预措施至关重要。即使我拥有博士学位，也了解 Ackerman 的研究，但我以前从未真正想到，我们可以用这种方式处理可解释人工智能的问题。”（P15，用户体验研究员）

揭示了现有组织实践中的盲点。参与者表示，“这个框架帮助 [他们] 清晰地指出了盲点”（P19）。例如，在获取价值观模块的业务单元时，我们发现，问责机制的构建方式与分析师感到脆弱之间存在价值冲突，这影响了他们的可操作性并损害了参与度（根本问题）。最重要的是，“这个框架帮助 [利益相关者] 认识到，[他们] 需要四个W——也就是社会透明度，而技术透明度并不足以解决问题”（P6）。参与者认为，“如果 [他们] 没有经历这一过程，这些盲点可能就无法被发现和解决”（P21）。

协作过程改善了部门间的互动。“经历这些区块带来的最佳成效之一，是改善了 [参与者] 与其他部门的关系。在 SecureCorp 这样的大型组织中，工作往往按职能条线彼此割裂。参与者分享说，这个框架“拓展了外围视野，让大家看到还有谁参与其中、他们能提供什么帮助；最重要的是，它让所有人齐聚一堂，共同解决这一问题”（P17）。尽管“一开始并不顺利， [非工程领域的参与者] 与工程团队意见不合，但 [他们] 如今建立了更好的工作关系，也增进了相互理解”（P15）。参与者表示，这一过程“让来自不同圈子的人走到了一起，这在一个复杂组织中是一项了不起的成就”（P22）。

这一过程将潜在的、既有的组织信息转化为可操作知识。 对我们而言，能够重复利用并重新调整现有实践和指南，是争取认同并推动采纳的一大优势。正如 第3节所述，我们有意识地尽可能利用现有基础设施；例如，我们采用了熟悉的人机交互技术，以及模型卡等已内置采纳机制的人工智能指南。产品经理尤其赞赏这种“可持续的思维方式，因为它充分利用了

“ [公司] 已有的资源”（P9）。这些选择减轻了参与者的额外负担，因此受到他们的认可。这也在没有给利益相关者增加显著负担的情况下，将潜在信息转化为可操作知识。以下是这位参与者的补充说明：

> “感谢你们没有强迫我们再做一份新的检查清单。研究人员常常让我们做新的事情，却不考虑我们如何利用已有的东西，这让我们感到厌倦。我很高兴我们充分利用了现有的信息和流程。我很喜欢这个过程帮助我们以新的视角看待旧事物，并解决了一个棘手的问题。”（P17，技术副总裁）

除了使用该框架带来的益处，参与者还指出了两项主要挑战。第一项挑战是“每个区块的所有权和责任归属问题”（P1）。参与者表示，他们原本期待框架能提供更多指导。来自负责任人工智能和政策团队的人员提出疑问：“既然这是一个多利益相关者问题，就不清楚哪个区块由谁负责。最终由谁承担问责？如果没有明确的指导，我们可能会互相推卸责任，或抱着‘这不关我的事’的想法”（20）。当被问及如何缓解这一障碍时，我们收到了富有启发性且具有建设性的建议。首先，鉴于问题情境可能各不相同，参与者一致认为，“不可能预先决定 [所有权]，因此不应由上而下地决定”（P22）。应当“在 [利益相关者]清楚了解问题之后”再讨论责任归属（P4）。其次，由于框架中的每个区块通常涉及多个利益相关者，所有权“不必只归属于一个团队，也可以由多个团队共同承担”（P12）。可以通过询问“如果这个区块出了问题，哪个群体损失最大？”来确定该区块问责责任的相对分配。若出现分歧，“领导层可能需要主动介入”（P20）。最后，对话应尽早开始——“在建立基线理解之前就开始，并在更新后的理解阶段重新讨论，以确认情况是否发生变化”（P2）。

另一个挑战在于框架早期阶段**难以获得集体认同**。在这一阶段，团队难以获得认同，希望框架能更好地引导这一过程。数据科学家和开发人员“起初难以理解，为什么讨论模型时需要让非开发人员参与进来”（P7）。参与者为这一认同难题提出了切实可行的解决方案。他们建议“在入职培训期间与团队负责人召开全员会议。要获得认同，必须清楚地说明价值主张——对每个团队来说有什么好处？如果大家不齐心协力，我们又会失去什么？”（P18）。为了让利益相关者产生共同的主人翁意识，从而促进认同，参与者倡导并认可我们的参与式方法。他们“感到自己参与其中；显然， [他们] 是在与 [研究人员]一起解决问题”（P16）。这位产品经理一语中的：“如果你想让大家加入，首先要让他们觉得自己的声音很重要；接着，要尽早清楚地说明，这是一项团队协作，大家都需要尽一份力”（P9）

6 讨论与启示

6.1 将框架迁移到新领域

我们秉持可定制性与多功能性的理念构建这一框架，力求在具体性与普遍性之间取得平衡。因此，我们从不同领域着手开发该框架，并将其应用于第三个全新的领域。领域的多样性是一项优势，也体现了该框架应用的多功能性。尽管这些领域各不相同，却有一个重要的共同点：所有领域都存在社会技术鸿沟。因此，该框架也可应用于销售和网络安全之外的领域，例如机器人学、自动驾驶等。只要问题涉及嵌入社会技术环境中的可解释性，就可以使用该框架来描绘社会技术鸿沟。这是因为我们所提出的问题、对每个区块的操作化方式以及入门问题都相似。此外，将该框架迁移到新领域时，

我们应留意相关情境是否适用。可迁移性本身就包含情境敏感性 [53, 70, 144]。现实世界中的可解释人工智能设置具有不同的情境、领域和应用范围，因此，我们的框架能否迁移到新领域取决于情境敏感性。鉴于该框架的构建方式，以及所有社会技术系统中都存在社会技术鸿沟，我们的框架可能具有广泛的可迁移性，但并非普遍适用。

在将框架调整到新的领域时，我们提供三个可供考虑的“旋钮”（实体）：合作的性质、基础设施的复杂程度，以及利害关系（或后果）。这三个旋钮并非可能影响框架迁移的全部因素，也并非旨在穷尽所有因素。不过，它们很有参考价值，可在迁移过程中提供指导。

对于合作的性质，工作的合作程度越高，框架社会维度的影响就越大，社会透明度在弥合社会因素与技术因素方面的潜在作用也越大。我们的使用场景中合作程度相对较高——在这三个案例中，都有多位用户与人工智能系统进行交互。相比之下，如果使用场景针对的是一对一的人机交互，那么社会因素的权重可能就没那么大。除了参与人机交互的用户数量外，我们还应考虑地理共址的性质和组织文化。员工分布越广泛（例如跨国企业），可能越需要较高水平的社会透明度。

对于基础设施的复杂性，我们应同时考虑技术层面和社会（组织）层面的复杂性。技术复杂性可能包括复杂的数据基础设施、模型（深度神经网络集成）等。在社会层面，复杂的组织结构会影响我们如何理解价值观之间的相互作用、可操作性的实现方式，以及哪些人为因素影响对人工智能的信任。复杂程度可能与我们需要对每个区块进行的“遍历”或迭代次数相关——复杂程度越高，我们可能就需要投入更多时间和精力来落实每个区块（例如，为集成模型编制多份模型卡），而为了更新理解而进行多次迭代的可能性也越大。因此，我们的框架提供了主题标签（基线理解、更新后的理解等），以便管理进度并跟踪挑战。

对于利害程度（或后果），我们应谨记其最终对人类生活的影响。这些影响可能涉及分配正义问题，例如（通过算法）分配救济资金；也可能是用户因接受人工智能的假阳性决策而面临失业风险。负责推荐音乐的人工智能系统，其利害程度与负责推荐贷款审批的系统不同。例如，在 Atlas（网络安全）案例研究中，利害程度很高——一个步骤的疏漏就可能让某人丢掉工作（Julie 的故事就是一个例子）。人工智能的错误所带来的责任完全落在了人类身上。这种问责的转移造成了敏感局面，需要谨慎进行利益相关方管理。我们不仅必须重新审视社会维度中的区块（例如可操作性），以更新理解；还必须谨慎落实价值观区块——与领导层沟通时，我们必须把握好平衡，避免让任何一方觉得自己受到指责。因此，考虑利害程度时，还应考虑社会政治层面。

6.2 实践意义

我们的工作具有实践意义，有助于研究人员和从业人员（技术开发端的利益相关者，例如数据科学家、人工智能工程师、技术产品经理、用户体验研究人员）将我们提出的框架付诸实践。

6.2.1 对从业人员的启示。首先，以社会技术视角思考会带来切实收益——组织环境中的技术问题都是社会技术问题。这意味着它们既包含社会组成部分，也包含技术组成部分。因此，单靠技术干预无法解决社会技术问题。在寻找解决方案时，开发人员和数据科学家若能

同时考虑社会和技术维度，就可能取得进展。其次，解决问题需要多利益相关者参与；因此，从业人员应主动寻求与其他团队合作。正如我们在框架有效性研究（第5.7节）中所强调的，工程师和开发人员得知自己并非独自应对人工智能中的社会技术复杂问题后，感到如释重负。该框架帮助他们意识到，“这是一场多人参与的游戏”（P2），而纯粹的技术视角会给工程师带来不公平的负担，让他们去解决涉及多利益相关者且具有社会技术性质的问题。第三，即使社会技术鸿沟可能永远无法完全弥合，开展绘制鸿沟这一活动仍有其价值。正如我们在第5节中所看到的，从揭示组织层面的价值冲突，到发掘不易察觉的技术局限，由框架引导的流程不仅为理解可解释人工智能中的社会技术鸿沟提供了可操作的洞见，也说明了如何应对这一鸿沟。该流程还带来了一个积极的附带效果：揭示了可能妨碍跨职能团队合作的组织障碍；若能通过正式或非正式方式加以解决，便可改善未来的最佳实践。

6.2.2 对研究人员的启示。首先，采用参与式方法有助于获得利益相关者的认同——如果利益相关者觉得自己是在与研究人员共同设计，而不是由研究人员为他们设计，他们往往会觉得自己受到重视，并积极投入、主动承担整个过程。我们将所有问题都界定为合作解决的共同问题 [1]，并尽可能平衡其中的权力动态。参与式方法也可能使集体认同变得复杂（这是我们在有效性研究中发现的一项挑战）；例如，谁来决定某个区块的事项。我们在第5节分享了缓解策略，例如在启动流程前清晰阐明价值主张。其次，研究人员在引入新的检查清单和流程之前，应充分利用现有信息和文档。正如我们的框架有效性研讨会所展示的（见第5.7节），这种可持续的方法提高了（a）获得利益相关者认同和接受的可能性，（b）将潜在信息转化为可操作知识，同时不显著增加利益相关者的负担，以及（c）缓解了参与者所说的检查清单倦怠，即人们因冗长的检查清单而感到不堪重负、参与度下降。这也是我们的框架鼓励使用成熟的检查清单指南和熟悉的实验方法的原因。第三，在可行的情况下，我们建议在参与门槛较低的小组环境中开展活动，因为这类环境往往能激发更多想法。诸如“午餐学习会”这类相对低压力的小组活动，员工可以聚在一起聆听演讲，是激发人们对项目初步兴趣的有效途径。相关利益相关者可能已经在场，因此更容易把大家聚到同一空间。这也是我们尽可能开展研讨会的原因。我们经常通过午餐学习会招募参与者，并利用虚拟研讨会降低参与门槛。我们希望这些实践启示能帮助未来的研究人员和从业人员在现实环境中将该框架付诸实践。

6.3 概念与理论意义

通过将社会技术鸿沟置于可解释人工智能的情境中，我们的框架以三种方式拓展了概念版图，从而丰富了其认识论。首先，它展现了从概念层面理解鸿沟的价值，将其与实际成果联系起来。通过提供一种系统化、由流程驱动的结构来描绘这一鸿沟，我们的框架展示了理解鸿沟这一行为本身如何催生旨在弥合鸿沟的设计干预。这种联系凸显了在实践中追求Ackerman提出鸿沟概念时的首要目标——理解鸿沟的必要性 [3]。这也体现了我们在引言中倡导的重点转变：在应对可解释人工智能中的社会技术鸿沟时，从“填补鸿沟”的理念转向“理解鸿沟”的理念。

其次，我们的框架串联了人工智能和人机交互领域的不同研究脉络，并将其转化为可解释人工智能情境下的内容，从而丰富其概念版图。例如，在技术侧，我们整合了

现有指南（如数据集数据表 [59], 模型卡 [109], XAI 问题库 [87],等），这些指南此前各自存在于不同的领域。在社会侧，我们从可解释人工智能范式（如社会透明度 [43]）中提炼出转化性见解，这些范式纳入了社会组织层面的要素。更重要的是，该框架阐明了各侧构成模块之间的相互作用，并提供了方法论指南（如价值敏感设计（VSD） [58] 和参与式设计（PD） [136]），以指导目标的实际落实。

第三，通过划定边界，将社会因素而非纯粹的技术因素纳入其中，我们的框架重新构想了相关因素，并有可能赋能利益相关者避免陷入某些技术中心主义陷阱，例如解决方案主义 [114] 和形式主义 [63]。Selbst 等人[137] 指出，陷入这些陷阱（排除社会情境）可能导致人工智能干预加剧社会不平等和不公平。通过“从过程而非解决方案的角度重新聚焦设计” [137], ，我们的框架可以帮助研究人员和从业人员避免这些陷阱。它鼓励采用一种以社会技术视角为基础的“异质性工程方法” [137],，批判性地思考应采用哪些社会因素、为何采用，以及何时实施设计干预 [63, 137]。例如，在 Atlas 案例研究中，该框架不仅帮助我们围绕社会透明度 [43], 的四个W（谁、什么、何时、为何）制定设计建议，还指出了其影响最显著的领域（可操作性和信任）。

概念层面的启示与我们的 设计哲学密切相关。正如我们在引言中所述，我们的框架具有生成性，而非规范性。其以流程为驱动的特性，旨在赋能利益相关者（例如研究人员、设计人员和开发人员），使其能够系统地绘制社会技术鸿沟，并生成解决这一鸿沟的见解。我们不认为存在一种能够普遍适用于所有情境、绘制鸿沟的一刀切范式。通往同一目的地的路径可能有很多。因此，我们不对映射鸿沟的具体路径施加限制。相反，我们着重于系统地将各个构成模块付诸实践，由生成性见解引导寻路过程。例如，与大型跨国组织相比，初创公司在绘制鸿沟时可能采取不同的行动顺序。尽管它们面临的鸿沟类型可能不同，但鸿沟依然存在，而我们的生成性框架可以在这两种情况下发挥作用。融入由Pinch 和 Bijker 提出的技术的社会建构理论 [118], 后，我们鼓励在流程中保持解释灵活性，并从不同的相关社会群体出发进行考量，每个群体对技术都有自己的目标和愿景。这种设计哲学不仅贯穿于整个框架，也对未来基于该框架开展的工作具有概念层面的启示。

6.4 局限性与未来工作

我们迈出了形成性的一步：通过两个案例研究推导出一个分析框架，并将其应用于另一个案例，从而描绘差距并设法弥合差距。鉴于这是初步尝试，我们的研究所得见解应据此限定适用范围。未来，我们需要在更多领域和不同情境中应用该框架。在我们的案例研究中，开发团队和终端用户都属于同一个组织。未来的工作可以探索开发团队与终端用户团队分属不同组织的情境（例如，服务提供商与客户）。

在未来的迭代中，我们受到 Agre 的批判性技术实践设计哲学的启发 [5]。与 Agre 的观点一致，“至少在可预见的未来，[我们] 需要兼具双重身份——一只脚扎根于设计的实践工作，另一只脚扎根于批判性的反思工作。” [6]。因此，至少在可预见的未来，我们对可解释人工智能中社会技术鸿沟的探索仍在进行中。我们通过提供一个有助于映射并弥合这一鸿沟的框架，在设计工作中“迈出了一只脚”。现在，我们希望与更广泛的人机交互和可解释人工智能社区共同学习，在批判性的自我反思领域“迈出另一只脚”。

7 结论

可解释人工智能（可解释人工智能）系统本质上是社会技术系统，因此受到社会技术鸿沟的影响。本文聚焦于绘制可解释人工智能系统中技术可供性与社会需求之间社会技术鸿沟这一问题。我们在两个不同领域（销售和心理健康）采用了两个案例研究，以实证推导出一个分析框架，帮助绘制可解释人工智能中的社会技术鸿沟。该框架由技术侧和社会侧各三个构成模块组成。技术侧包括数据、模型和解释三个组成部分，社会侧则包括信任、可操作性和价值观三个组成部分。针对每个构成模块，我们都提供了一系列入门问题，以帮助弥合可解释人工智能系统中的社会技术鸿沟。此外，我们还从缓解措施和一阶近似的角度，阐述了如何弥合这一鸿沟 [3]。接下来，我们将所提出的框架应用于一个新领域（网络安全）的第三个案例研究，以展示其可供性，以及它如何不仅帮助理解问题，还能生成可操作的洞见，从而妥善解决问题并提升系统的可解释性。我们的工作提出了设计建议，并指出了映射和弥合可解释人工智能中社会技术鸿沟时可能遇到的挑战、风险和张力。

致谢

我们衷心感谢参与者为本项目慷慨投入时间。没有他们的意见，本项目便不可能完成。我们也要感谢各组织以及案例研究的开展地点，感谢他们给予的合作。我们感谢佐治亚理工学院以人为中心的人工智能实验室和社会动态与福祉实验室的成员，他们持续提供意见，使本文提出的概念更加完善。我们由衷感谢 Vera Liao、Michael Muller 和 Samir Passi 提供的宝贵反馈，帮助我们恰当地界定项目范围。特别感谢 Rachel Ehsan 慷慨提供校对意见。本项目部分由美国国家科学基金会通过第 1928586 号资助项目提供支持。

参考文献

[1] Leena Aarikka-Stenroos 和 Elina Jaakkola。2012年。知识密集型商业服务中的价值共创：关于联合问题解决过程的二元视角。《工业营销管理》41, 1 (2012), 15–26。[2] Ashraf Abdul、Jo Vermeulen、Danding Wang、Brian Y Lim 和 Mohan Kankanhalli。2018年。可解释、可问责且可理解系统的趋势与发展轨迹：一项人机交互研究议程。载于*2018*年计算机系统中的人为因素会议论文集。ACM，582。[3] Mark S Ackerman。2000年。计算机支持的协同工作的智识挑战：社会需求与技术可行性之间的差距。《人机交互》(2000)。[4] Amina Adadi 和 Mohammed Berrada。2018年。窥探黑箱内部：可解释人工智能（XAI）综述。《*IEEE*开放获取期刊》6 (2018), 52138–52160。[5] P Agre。1997年。走向批判性技术实践：在 Bowker 探索改革人工智能的经验教训。载于《社会科学、技术系统与协作工作：超越巨大鸿沟》(1997)。[6] Philip E Agre。1997年。《计算与人类经验》。剑桥大学出版社。[7] Ahmed Alqaraawi、Martin Schuessler、Philipp Weiß、Enrico Costanza 和 Nadia Berthouze。2020年。评估卷积神经网络的显著性图解释：一项用户研究。载于第*25*届智能用户界面国际会议论文集。[8] Steven Alter。2010年。社会技术系统的设计空间。(2010)。[9] Saleema Amershi、Dan Weld、Mihaela Vorvoreanu、Adam Fourney、Besmira Nushi、Penny Collisson、Jina Suh、Shamsi Iqbal、Paul N Bennett、Kori Inkpen 等。2019年。人机交互指南。载于*2019*年计算机系统中的人为因素会议论文集。1–13。[10] McKane Andrus、Sarah Dean、Thomas Krendl Gilbert、Nathan Lambert 和 Tom Zick。2020年。面向公共利益的人工智能开发：从抽象陷阱到社会技术风险。载于*2020*年*IEEE*技术与社会国际研讨会（*ISTAS*）。电气电子工程师学会。[11] Matthew Arnold、Rachel KE Bellamy、Michael Hind、Stephanie Houde、Sameep Mehta、Aleksandra Mojsilović、Ravi Nair、K Natesan Ramamurthy、Alexandra Olteanu、David Piorkowski 等。2019年。FactSheets：通过供应商符合性声明提升对人工智能的信任

服务。IBM《研究与开发杂志》63, 4/5 (2019), 6–1。[12] Alejandro Barredo Arrieta、Natalia Díaz-Rodríguez、Javier Del Ser、Adrien Bennetot、Siham Tabik、Alberto Barbado、Salvador García、Sergio Gil-López、Daniel Molina、Richard Benjamins 等。2020年。可解释人工智能（XAI）：概念、分类体系、机遇与挑战，以及迈向负责任人工智能的路径。《信息融合》58 (2020)。[13] Vijay Arya、Rachel KE Bellamy、Pin-Yu Chen、Amit Dhurandhar、Michael Hind、Samuel C Hoffman、Stephanie Houde、Q Vera Liao、Ronny Luss、Aleksandra Mojsilović 等。2019年。一种解释并非适合所有人：人工智能可解释性技术工具包与分类体系。arXiv预印本 arXiv:1909.03012 (2019)。[14] Maryam Ashoori 和 Justin D Weisz。2019年。我们信任人工智能吗？影响人工智能赋能决策过程可信度的因素。arXiv预印本 arXiv:1912.02675 (2019)。[15] 美国精神医学学会等。2015年。《美国精神医学学会成人精神科评估实践指南》。美国精神医学学会出版社。[16] Aaron Balick。2014年。技术、社交媒体与心理治疗：跟上时代步伐。《当代心理治疗》6, 2 (2014)。[17] Natalya N Bazarova、Yoon Hyung Choi、Victoria Schwanda Sosik、Dan Cosley 和 Janis Whitlock。2015年。Facebook上的情绪社交分享：渠道差异、满意度与回复。载于第18届ACM计算机支持的协同工作与社会计算会议论文集。154–164。[18] GM Beal 和 JM Bohlen。1957年。扩散过程（特别报告第18号，农业试验站）。爱荷华州立学院 (1957)。[19] Martin Bella 和 Bruce Hanington。2012年。《通用设计方法》。马萨诸塞州贝弗利：Rockport 出版社 (2012), 204。[20] Rachel KE Bellamy、Kuntal Dey、Michael Hind、Samuel C Hoffman、Stephanie Houde、Kalapriya Kannan、Pranay Lohia、Jacquelyn Martino、Sameep Mehta、Aleksandra Mojsilović 等。2018年。AI Fairness 360：用于检测、理解和缓解非预期算法偏见的可扩展工具包。arXiv预印本 arXiv:1810.01943 (2018)。[21] Hugh R Beyer 和 Karen Holtzblatt。1996年。情境技术入门套件。《Interactions》3, 6 (1996), 44–50。[22] Virginia Braun 和 Victoria Clarke。2006 3, 2 2006 77 101。心理学中的主题分析。《心理学定性研究》( ), –。[23] Zana Buçinca、Phoebe Lin、Krzysztof Z Gajos 和 Elena L Glassman。2020年。在评估可解释人工智能系统时，代理任务和主观测量指标可能产生误导。载于第25届智能用户界面国际会议论文集。[24] Joy Buolamwini 和 Timnit Gebru。2018年。性别阴影：商业性别分类中的交叉性准确率差异。载于公平性、问责制与透明度会议。机器学习研究论文集，77–91。[25] Daniel Buschek、Lukas Mecke、Florian Lehmann 和 Hai Dang。2021年。设计人机协同创作系统时的九个潜在陷阱。arXiv预印本 arXiv:2104.00358 (2021)。[26] Carrie J Cai、Jonas Jongejan 和 Jess Holbrook。2019年。基于示例的解释对机器学习界面的影响。载于第24届智能用户界面国际会议论文集。[27] Diogo V Carvalho、Eduardo M Pereira 和 Jaime S Cardoso。2019年。机器学习可解释性：方法与指标综述。《电子学》8, 8 (2019), 832。[28] Stevie Chancellor 和 Munmun De Choudhury。2020年。社交媒体心理健康状况预测技术的方法：批判性综述。《NPJ数字医学》(2020)。[29] Zhengping Che、Sanjay Purushotham、Robinder Khemani 和 Yan Liu。2016年。用于重症监护病房结局预测的可解释深度模型。载于《美国医学信息学会年度研讨会论文集》，第2016卷。[30] Hao-Fei Cheng、Ruotong Wang、Zheng Zhang、Fiona O'Connell、Terrance Gray、F Maxwell Harper 和 Haiyi Zhu。2019年。通过用户界面解释决策算法：帮助非专家利益相关者的策略。载于2019年计算机系统中的人为因素会议论文集。[31] EunJeong Cheon 和 Norman Makoto Su。2016年。将机器人学家价值观融入面向人形机器人的价值敏感设计框架。载于2016年第11届ACM/IEEE人机交互国际会议。电气电子工程师学会，375–382。[32] EunJeong Cheon 和 Norman Makoto Su。2018年。未来主义自传：编织参与者叙事以引出关于机器人的价值观。载于2018年ACM/IEEE人机交互国际会议论文集。[33] Clifford Christians。1989年。规范性技术理论。载于《技术变革》。施普林格，123–139。[34] Massimiliano Sassoli de Bianchi。2013年。观察者效应。《科学基础》18, 2 (2013), 213–243。[35] Munmun De Choudhury、Scott Counts 和 Eric Horvitz。2013年。通过社交媒体预测产后情绪与行为变化。载于SIGCHI计算机系统中的人为因素会议论文集（CHI '13）。[36] Munmun De Choudhury、Scott Counts、Eric J. Horvitz 和 Aaron Hoff。2014年。根据共享的 Facebook 数据描述并预测产后抑郁症。载于第17届ACM计算机支持的协同工作与社会计算会议论文集（CSCW '14）。美国计算机协会出版社，美国马里兰州巴尔的摩，626–638。[37] Munmun De Choudhury、Min Kyung Lee、Haiyi Zhu 和 David A Shamma。2020年。本特刊导言：统一人机交互与人工智能。《人机交互》(2020)。

[38] Shipi Dhanorkar、Christine T Wolf、Kun Qian、Anbang Xu、Lucian Popa 和 Yunyao Li。2021年。谁需要在何时了解什么？：通过考察整个人工智能生命周期中的解释，拓展可解释人工智能（可解释人工智能）的设计空间。载于《交互系统设计大会 2021》。[39] Jonathan Dodge、Q Vera Liao、Yunfeng Zhang、Rachel KE Bellamy 和 Casey Dugan。2019年。解释模型：解释如何影响公平性判断的实证研究。载于《智能用户界面国际会议》第24届论文集。[40] Finale Doshi-Velez 和 Been Kim。2017年。迈向可解释机器学习的严谨科学。arXiv预印本 arXiv:1702.08608（2017年）。[41] Paul Dourish 和 Genevieve Bell。2011年。预见数字未来：普适计算中的混乱与神话。麻省理工学院出版社。[42] Graham Dove、Kim Halskov、Jodi Forlizzi 和 John Zimmerman。2017年。用户体验设计创新：将机器学习作为设计材料时面临的挑战。《2017年计算机系统中的人为因素会议论文集》- CHI ’ *17* 2017 278 288 10 1145 3025453 3025739 ( ), –。https://doi.org/ 。/ 。Upol [43] Ehsan、Q Vera Liao、Michael Muller、Mark O Riedl 和 Justin D Weisz。2021年。扩展可解释性：迈向人工智能系统中的社会透明度。载于《2021年计算机系统中的人为因素会议论文集》。[44] Upol Ehsan、Samir Passi、Q Vera Liao、Larry Chan、I Lee、Michael Muller、Mark O Riedl 等。2021年。可解释人工智能中的“谁”：人工智能背景如何塑造人们对人工智能解释的认知。arXiv预印本 arXiv:2107.13509（2021年）。[45] Upol Ehsan 和 Mark O Riedl。2020年。以人为中心的可解释人工智能：迈向反思性的社会技术方法。载于《人机交互国际会议》。施普林格，449–466。[46] Upol Ehsan 和 Mark O Riedl。2021年。可解释性陷阱：超越可解释人工智能中的暗黑模式。arXiv预印本 arXiv:2109.12480（2021年）。[47] Upol Ehsan 和 Mark O Riedl。2022年。XAI的社会建构：我们是否需要一个能够统摄一切的定义？arXiv预印本 arXiv:2211.06499（2022年）。[48] Upol Ehsan、Pradyumna Tambwekar、Larry Chan、Brent Harrison 和 Mark Riedl。2019年。自动化理由生成：一种用于可解释人工智能的技术及其对人类感知的影响。载于《智能用户界面国际会议》论文集。[49] Upol Ehsan、Philipp Wintersberger、Q Vera Liao、Elizabeth Anne Watkins、Carina Manger、Hal Daumé III、Andreas Riener 和 Mark O Riedl。2022年。以人为中心的可解释人工智能（以人为中心的可解释人工智能）：超越打开人工智能的黑箱。载于《CHI计算机系统中的人为因素会议扩展摘要》。[50] Malin Eiband、Daniel Buschek 和 Heinrich Hussmann。2021年。如何支持用户理解智能系统？构建讨论框架。载于《智能用户界面国际会议》第26届。120–132。[51] Malin Eiband、Hanna Schneider、Mark Bilandzic、Julian Fazekas-Con、Mareike Haug 和 Heinrich Hussmann。2018年。将透明性设计付诸实践。载于《智能用户界面国际会议》第23届。211–223。[52] Motahhare Eslami、Aimee Rickman、Kristen Vaccaro、Amirhossein Aleyasen、Andy Vuong、Karrie Karahalios、Kevin Hamilton 和 Christian Sandvig。2015年。“我一直以为我和 [她]并没有那么亲近。”对新闻信息流中不可见算法的推理。载于《CHI会议：计算机系统中的人为因素》论文集。[53] Deborah Finfgeld-Connett。2010年。元综合研究成果的可推广性与可迁移性。《高级护理杂志》66，2（2010年），246–254。[54] Carl E. Fisher 和 Paul S. Appelbaum。2017年。超越谷歌搜索。《哈佛精神病学评论》（2017年），1。https: //doi.org/ 10 1097 0000000000000145 。/hrp。Andrea Forte 和 Cliff Lampe。 。定义、理解并支持开放协作：文献中的经验。《美国行为科学家》57，5（2013年），535–547。[55] 2013[56] Batya Friedman。1996年。价值敏感设计。《Interactions》3，6（1996年），16–23。[57] Batya Friedman 和 David Hendry。2012年。构想卡片：激发人文与技术想象力的工具包。载于《SIGCHI会议：计算机系统中的人为因素》论文集。1145–1148。[58] Batya Friedman、Peter Kahn 和 Alan Borning。2002年。价值敏感设计：理论与方法。华盛顿大学技术报告 2-12（2002年）。[59] Timnit Gebru、Jamie Morgenstern、Briana Vecchione、Jennifer Wortman Vaughan、Hanna Wallach、Hal Daumé III 和 Kate Crawford。2018年。数据集数据表。arXiv预印本 arXiv:1803.09010（2018年）。[60] Katy Ilonka Gero、Zahra Ashktorab、Casey Dugan、Qian Pan、James Johnson、Werner Geyer、Maria Ruiz、Sarah Miller、David R Millen、Murray Campbell 等。2020年。合作博弈环境中的人工智能代理心智模型。载于《2020年计算机系统中的人为因素会议论文集》。1–12。[61] Bhavya Ghai、Q Vera Liao、Yunfeng Zhang、Rachel Bellamy 和 Klaus Mueller。2021年。可解释主动学习（XAL）：将人工智能解释作为机器教师的界面。ACM人机交互会议论文集 4，CSCW3（2021年）。[62] Leilani H Gilpin、David Bau、Ben Z Yuan、Ayesha Bajwa、Michael Specter 和 Lalana Kagal。2018年。解释解释：一种评估机器学习可解释性的方法。arXiv预印本 arXiv:1806.00069（2018年）。

[63] Ben Green 和 Salomé Viljoen。2020年。算法现实主义：拓展算法思维的边界。载于《2020年公平性、问责制与透明度会议》论文集。19–31。[64] Riccardo Guidotti、Anna Monreale、Salvatore Ruggieri、Dino Pedreschi、Franco Turini 和 Fosca Giannotti。2018年。黑箱决策系统的局部基于规则的解释。arXiv预印本 arXiv:1805.10820（2018年）。[65] Riccardo Guidotti、Anna Monreale、Salvatore Ruggieri、Franco Turini、Fosca Giannotti 和 Dino Pedreschi。2018年。解释黑箱模型的方法综述。《ACM计算调查》（CSUR）51，5（2018年），1–42。[66] David Gunning。2017年。可解释人工智能（可解释人工智能）。美国国防高级研究计划局（DARPA），未注明日期，Web 2（2017年），2。[67] Sharath Chandra Guntuku、David B Yaden、Margaret L Kern、Lyle H Ungar 和 Johannes C Eichstaedt。2017年。在社交媒体上检测抑郁症和精神疾病：综合性综述。《行为科学当前观点》（2017年）。[68] Karen Hao。2019年。人工智能正在把人送进监狱——而且还判断错了。《麻省理工科技评论》（2019年1月21日）。检索于2019年8月26日，网址：https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/ [69] MD Romael Haque、Katherine Weathington、Joseph Chudzik 和 Shion Guha。2020年。理解执法部门和普通民众对设计可解释犯罪地图算法的观点。载于《2020年计算机支持的协同工作与社会计算会议伴随出版物》。269–273。[70] Gillian R Hayes。2011年。行动研究与人机交互的关系。《ACM人机交互汇刊》（TOCHI）18，3（2011年），1–20。[71] Michael Hind。2019年。解释可解释人工智能。《XRDS：十字路口》，ACM学生杂志 25，3（2019年），16–19。[72] Michael Hind、Sameep Mehta、Aleksandra Mojsilovic、Ravi Nair、Karthikeyan Natesan Ramamurthy、Alexandra Olteanu 和 Kush R Varshney。2018年。通过供应商符合性声明提高对人工智能服务的信任。arXiv预印本 arXiv:1808.07261 18（2018年），2813–2869。[73] Kevin Anthony Hoff 和 Masooda Bashir。2015年。对自动化的信任：整合影响信任因素的实证证据。《人为因素》57，3（2015年），407–434。[74] Fred Hohman、Andrew Head、Rich Caruana、Robert DeLine 和 Steven M Drucker。2019年。Gamut：一种理解数据科学家如何理解机器学习模型的设计探针。载于《2019年CHI会议：计算机系统中的人为因素》论文集。[75] Sarah Holland、Ahmed Hosny 和 Sarah Newman。2020年。数据集营养标签。《数据保护与隐私：数据保护与民主》（2020年）1（2020年）。[76] Andreas Holzinger、Chris Biemann、Constantinos S Pattichis 和 Douglas B Kell。2017年。我们需要什么来构建面向医疗领域的可解释人工智能系统？arXiv预印本 arXiv:1712.09923（2017年）。[77] Andrew JI Jones、Alexander Artikis 和 Jeremy Pitt。2013年。智能社会技术系统的设计。《人工智能评论》39，1（2013年），5–20。[78] Gajendra Jung Katuwal 和 Robert Chen。2016年。面向精准医学的机器学习模型可解释性。arXiv预印本 arXiv:1610.09045（2016年）。[79] Harmanpreet Kaur、Harsha Nori、Samuel Jenkins、Rich Caruana、Hanna Wallach 和 Jennifer Wortman Vaughan。2020年。解读可解释性：理解数据科学家对机器学习可解释性工具的使用。载于《2020年计算机系统中的人为因素会议论文集》。[80] Jon Kleinberg、Himabindu Lakkaraju、Jure Leskovec、Jens Ludwig 和 Sendhil Mullainathan。2017年。人类决策与机器预测。《经济学季刊》 133, 1 2017 237 293 10 1093 032 ( )，–。[81] https://doi.org/ 。/qje/qjx Bran Knowles 和 John T Richards。 。2021 权威的制裁：促进公众对人工智能的信任。载于《2021年ACM公平性、问责制与透明度会议》论文集。262–271。[82] Sivam Krish。2011年。一种实用的生成式设计方法。《计算机辅助设计》43，1（2011年），88–100。[83] Todd Kulesza、Simone Stumpf、Margaret Burnett、Sherry Yang、Irwin Kwan 和 Weng-Keen Wong。2013年。太多、太少，还是恰到好处？解释影响终端用户心智模型的方式。载于《2013年IEEE可视化语言与以人为中心的计算研讨会》。IEEE。[84] Vivian Lai、Han Liu 和 Chenhao Tan。2020年。“为什么‘芝加哥’具有欺骗性？”迈向构建面向人类的模型驱动教程。载于《2020年计算机系统中的人为因素会议论文集》。[85] Ellen E Lee、John Torous、Munmun De Choudhury、Colin A Depp、Sarah A Graham、Ho-Cheol Kim、Martin P Paulus、John H Krystal 和 Dilip V Jeste。2021年。用于心理健康护理的人工智能：临床应用、障碍、促进因素与人工智能智慧。《生物精神病学：认知神经科学与神经成像》（2021年）。[86] Min Kyung Lee、Daniel Kusbit、Anson Kahng、Ji Tae Kim、Xinran Yuan、Allissa Chan、Daniel See、Ritesh Noothigattu、Siheon Lee、Alexandros Psomas 等。2019年。WeBuildAI：算法治理的参与式框架。ACM人机交互会议论文集 3，CSCW（2019年），1–35。[87] Q Vera Liao、Daniel Gruen 和 Sarah Miller。2020年。质疑人工智能：为可解释人工智能用户体验提供设计实践参考。载于《2020年计算机系统中的人为因素会议论文集》。[88] Q. Vera Liao、Milena Pribić、Jaesik Han、Sarah Miller 和 Daby Sow。2021年。面向可解释人工智能用户体验的问题驱动设计流程。arXiv:2104.*03483 [*cs*]*（2021年9月）。http://arxiv.org/abs/2104.03483 arXiv:2104.03483。

[89] Q Vera Liao 和 Kush R Varshney。2021年。以人为中心的可解释人工智能（xai）：从算法到用户体验。arXiv预印本 arXiv:2110.10790（2021年）。[90] Brian Y Lim 和 Anind K Dey。2009年。评估情境感知应用对可理解性的需求。载于第11届普适计算国际会议论文集。195–204。[91] Brian Y Lim 和 Anind K Dey。2010年。支持情境感知应用可理解性的工具包。载于第12届ACM普适计算国际会议论文集。13–22。[92] Brian Y Lim 和 Anind K Dey。2011年。探讨不确定情境感知应用的可理解性。载于第13届普适计算国际会议论文集。415–424。[93] Brian Y Lim、Anind K Dey 和 Daniel Avrahami。2009年。为何解释与为何不解释能够提升情境感知智能系统的可理解性。载于SIGCHI计算机系统中的人为因素会议论文集（CHI ’ *09* 2119 2128 10 1145 1518701 1519023 ）。ACM，纽约，美国，–。 2009 [94] https://doi.org/。/。Brian Y Lim、Anind K Dey 和 Daniel Avrahami。。为何解释与为何不解释能够提升情境感知智能系统的可理解性。载于SIGCHI计算机系统中的人为因素会议论文集。[95] Zachary C Lipton。2018年。模型可解释性神话。《Queue》16，3（2018年），31–57。[96] Tyler J. Loftus、Patrick J. Tighe、Amanda C. Filiberto、Philip A. Efron、Scott C. Brakenridge、Alicia M. Mohr、Parisa Rashidi、Jr Upchurch、Gilbert R. 和 Azra Bihorac。2020年。人工智能与外科决策。《JAMA Surgery》（2020年）。[97] Tania Lombrozo。2011年。解释的工具价值。《Philosophy Compass》6，8（2011年）。[98] Tania Lombrozo。2012年。解释与溯因推理。《牛津思维与推理手册》（2012年）。[99] Duri Long 和 Brian Magerko。2020年。什么是人工智能素养？能力与设计考量。载于2020年计算机系统中的人为因素会议论文集。1–16。[100] Ewa Luger 和 Abigail Sellen。2016年。“就像有个非常糟糕的私人助理”：用户对对话代理的期望与体验之间的鸿沟。载于2016年计算机系统中的人为因素会议论文集。5286–5297。[101] Henrietta Lyons、Eduardo Velloso 和 Tim Miller。2021年。概念化可争议性：对质疑算法决策的看法。《ACM人机交互会议论文集》5，CSCW1（2021年），1–25。[102] Donald MacKenzie。2018年 123，6 2018年 1635 1683 10 1086。《物质信号：高频交易的历史社会学》。《美国社会学杂志》（），697315 https://doi.org/。/ [103] Michael A Madaio、Luke Stark、Jennifer Wortman Vaughan 和 Hanna Wallach。2020年。共同设计检查清单，以了解人工智能公平性方面的组织挑战与机遇。载于2020年计算机系统中的人为因素会议论文集。1–14。[104] Jonathan Magnusson。 [n.d.]。提升终端用户的暗黑模式素养。（[n. d.]）。[105] Erin E Makarius、Debmalya Mukherjee、Joseph D Fox 和 Alexa K Fox。2020年。与机器共进：将人工智能引入组织的社会技术框架。《商业研究杂志》120（2020年）。[106] Masike Malatji、Sune Von Solms 和 Annlizé Marnewick。2019年。社会技术系统网络安全框架。《信息与计算机安全》（2019年）。[107] Jon McCormack、Alan Dorin 和 Troy Innocent。2004年。生成式设计：设计研究的一种范式。（2004年）。[108] Tim Miller。2019年。人工智能中的解释：来自社会科学的见解。《人工智能》267（2019年），1–38。[109] Margaret Mitchell、Simone Wu、Andrew Zaldivar、Parker Barnes、Lucy Vasserman、Ben Hutchinson、Elena Spitzer、Inioluwa Deborah Raji 和 Timnit Gebru。2019年。用于模型报告的模型卡。载于公平、问责与透明度会议论文集。220–229。[110] Brent Mittelstadt、Chris Russell 和 Sandra Wachter。2019年。解释人工智能中的解释。载于公平、问责与透明度会议论文集。279–288。[111] Shakir Mohamed、Marie-Therese Png 和 William Isaac。2020年。去殖民人工智能：将去殖民理论作为人工智能的社会技术前瞻。《哲学与技术》（2020年），1–26。[112] Sina Mohseni、Niloofar Zarei 和 Eric D Ragan。2018年。可解释人工智能系统设计与评估的多学科综述及框架。arXiv（2018年），arXiv–1811。[113] Geoffrey A Moore 和 Regis McKenna。1999年。跨越鸿沟。（1999年）。[114] Evgeny Morozov。2013年。想要拯救一切？点击此处：技术解决方案主义的谬误。Public Affairs。[115] Michael Muller 和 Q Vera Liao。 [n.d.]。通过参与式设计虚构探索人工智能伦理与价值观。（[n. d.]）。[116] Sean A Munson、Hasan Cavusoglu、Larry Frisch 和 Sidney Fels。2013年。利用社交媒体促进健康的社会技术挑战与进展。《医学互联网研究杂志》15，10（2013年），e226。[117] John Murawski。2019年。抵押贷款服务商借助人工智能加快住房贷款处理。《华尔街日报》（2019年3月18日）。检索于2020年9月16日，网址：https://www.wsj.com/articles/mortgage-providers-look-to-ai-to-process-home-loans-faster-11552899212

[118] Trevor J Pinch 和 Wiebe E Bijker。1984年。事实与人工制品的社会建构：或者，科学社会学与技术社会学如何相互受益。《科学社会研究》14，3（1984年），399–441。[119] Forough Poursabzi-Sangdeh、Daniel G Goldstein、Jake M Hofman、Jennifer Wortman Vaughan 和 Hanna Wallach。2018年。操纵和测量模型可解释性。arXiv预印本 arXiv:1802.07810（2018年）。[120] Mahima Pushkarna、Andrew Zaldivar 和 Oddur Kjartansson。2022年。数据卡：为负责任人工智能提供目的明确且透明的数据集文档。arXiv预印本 arXiv:2204.01075（2022年）。[121] Emilee Rader、Kelley Cotter 和 Janghee Cho。2018年。将解释作为支持算法透明度的机制。载于2018年计算机系统中的人为因素会议论文集。ACM，103。[122] Emilee Rader 和 Rebecca Gray。2015年。理解用户对Facebook动态消息中算法策展的看法。载于第33届ACM年度计算机系统中的人为因素会议论文集（CHI '15）。计算机协会，纽约，美国，

173 182 10 1145 2702123 2702174 –。

[123] 2018 https://doi.org/。/。Gabriëlle Ras、Marcel van Gerven 和 Pim Haselager。。深度学习中的解释方法：用户、价值观、关切与挑战。载于《计算机视觉与机器学习中的可解释和可诠释模型》。施普林格。[124] Marco Tulio Ribeiro、Sameer Singh 和 Carlos Guestrin。2016年。我为什么要信任你？：解释任意分类器的预测。载于第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集。ACM，1135–1144。[125] Marco Tulio Ribeiro、Sameer Singh 和 Carlos Guestrin。2018年。Anchors：高精度模型无关解释。载于AAAI人工智能会议论文集，第32卷。[126] Cynthia Rudin、Caroline Wang 和 Beau Coker。2020年。累犯预测中的保密与不公平时代。《哈佛数据科学评论》2，1（2020年3月31日）。https://doi.org/10.1162/99608f92.6ed64b30 https://hdsr.mitpress.mit.edu/pub/7z10o269。[127] Selma Šabanović。2010年。社会中的机器人，机器人中的社会。《国际社会机器人学期刊》2，4（2010年），439–450。[128] Koustuv Saha、Pranshu Gupta、Gloria Mark、Emre Kıcıman 和 Munmun De Choudhury。2023年。社交媒体使用中的观察者效应。（2023年）。[129] Koustuv Saha、Jordyn Seybolt、Stephen M Mattingly、Talayeh Aledavood、Chaitanya Konjeti、Gonzalo J Martinez、Ted Grover、Gloria Mark 和 Munmun De Choudhury。2021年。社交媒体上披露了哪些人生事件、以何种方式、何时以及由谁披露？载于2021年计算机系统中的人为因素会议论文集。1–22。[130] Koustuv Saha、Benjamin Sugar、John Torous、Bruno Abrahao、Emre Kıcıman 和 Munmun De Choudhury。2019年。一项关于使用精神科药物影响的社交媒体研究。载于AAAI网络与社交媒体国际会议论文集，第13卷。440–451。[131] Koustuv Saha、Asra Yousuf、Ryan L Boyd、James W Pennebaker 和 Munmun De Choudhury。2022年。社交媒体讨论可预测大学校园中的心理健康咨询。《科学报告》（2022年）。[132] Javier Sánchez-Monedero、Lina Dencik 和 Lilian Edwards。2020年。解决招聘歧视问题意味着什么？英国关于自动化招聘系统的社会、技术与法律观点。载于2020年公平性、问责制与透明度会议论文集。458–468。[133] Lindsay Sanneman 和 Julie A Shah。2020年。用于可解释人工智能设计与评估的基于情境感知的框架。载于《可解释、透明的自主智能体与多智能体系统》国际研讨会。[134] Benjamin Saunders、Julius Sim、Tom Kingstone、Shula Baker、Jackie Waterfield、Bernadette Bartlam、Heather Burroughs 和 Clare Jinks。2018年。定性研究中的饱和：探讨其概念化与操作化。《质量与数量》52，4（2018年），1893–1907。[135] Jakob Schoeffer 和 Niklas Kuehl。2021年。恰当的公平性感知？解释在帮助人们评估自动化决策系统公平性方面的有效性。载于2021年计算机支持的协同工作与社会计算会议伴随出版物。153–157。[136] Douglas Schuler 和 Aki Namioka。1993年。参与式设计：原则与实践。CRC出版社。[137] Andrew D Selbst、Danah Boyd、Sorelle A Friedler、Suresh Venkatasubramanian 和 Janet Vertesi。2019年。社会技术系统中的公平性与抽象。载于公平、问责与透明度会议论文集。[138] Phoebe Sengers、Kirsten Boehner、Shay David 和 Joseph'Jofish' Kaye。2005年。反思性设计。载于第4届十年一度的批判性计算会议：在理智与感性之间论文集。49–58。[139] Ben Shneiderman。2020年。以人为中心的人工智能：可靠、安全且值得信赖。《国际人机交互期刊》36，6（2020年），495–504。[140] Keng Siau 和 Weiyu Wang。2018年。在人工智能、机器学习和机器人学中建立信任。《Cutter商业技术期刊》31，2（2018年），47–53。[141] Supriya Singh、Anuja Cabraal、Catherine Demosthenous、Gunela Astbrink 和 Michele Furlong。2007年。密码共享：基于社会实践的安全设计启示。载于SIGCHI计算机系统中的人为因素会议论文集。895–904。[142] Alison Smith-Renner、Ron Fan、Melissa Birchfield、Tongshuang Wu、Jordan Boyd-Graber、Daniel S Weld 和 Leah Findlater。2020年。没有问责，就没有可解释性：一项关于解释与反馈的实证研究，

交互式机器学习。载于《2020年计算机系统中的人为因素会议论文集》。1–13。[143] Kacper Sokol 和 Peter Flach。2020年。可解释性事实表：系统评估可解释方法的框架。载于《2020年公平性、问责制与透明度会议论文集》。[144] Ernest T Stringer。2007年。行动研究（第三版）。（2007年）。[145] Simone Stumpf、Adrian Bussone 和 Dympna O'sullivan。2016年。解释是否有害？用户与机器学习系统的交互。载于 ACM SIGCHI以人为中心的机器学习研讨会。[146] Jiao Sun、Q Vera Liao、Michael Muller、Mayank Agarwal、Stephanie Houde、Kartik Talamadupula 和 Justin D Weisz。2022年。通过基于场景的设计探究生成式人工智能代码的可解释性。载于第27届智能用户界面国际会议。212–228。[147] Harini Suresh 和 John V Guttag。2019年。理解机器学习意外后果的框架。arXiv预印本 arXiv:1901.10002（2019年）。[148] Richard Tomsett、Dave Braines、Dan Harborne、Alun Preece 和 Supriyo Chakraborty。2018年。对谁而言可解释？用于分析可解释机器学习系统的基于角色的模型。arXiv预印本 arXiv:1806.07552（2018年）。[149] Jennifer Wortman Vaughan 和 Hanna Wallach。 [未注明日期]。1 以人为中心的可理解机器学习议程。（[未注明日期]）。[150] Viswanath Venkatesh 和 Fred D Davis。2000年。技术接受模型的理论扩展：四项纵向实地研究。《管理科学》46卷2期（2000年），186–204。[151] Viswanath Venkatesh、Michael G Morris、Gordon B Davis 和 Fred D Davis。2003年。用户对信息技术的接受：迈向统一观点。《MIS季刊》（2003年），425–478。[152] Daniel Vigo、Graham Thornicroft 和 Rifat Atun。2016年。估算精神疾病的真实全球负担。《柳叶刀·精神病学》3卷2期（2016年），171–178。[153] Guy H Walker、Neville A Stanton、Paul M Salmon 和 Daniel P Jenkins。2008年。社会技术系统理论综述：适用于新型指挥与控制范式的经典概念。《人体工程学科学理论问题》9卷6期（2008年）。[154] Danding Wang、Qian Yang、Ashraf Abdul 和 Brian Y Lim。2019年。设计理论驱动的以用户为中心的可解释人工智能。载于《2019年计算机系统中的人为因素会议论文集》。[155] Qiaosi Wang、Koustuv Saha、Eric Gregori、David Joyner 和 Ashok Goel。2021年。迈向人机交互中的相互心智理论：语言如何反映学生对虚拟教学助理的认知。载于《2021年计算机系统中的人为因素会议论文集》。1–14。[156] Daniel A Wilkenfeld 和 Tania Lombrozo。2015年。最佳解释推理（IBE）与为最佳推理而解释（EBI）。《科学与教育》24卷9-10期（2015年），1059–1077。[157] Christine Wolf 和 Jeanette Blomberg。2019年。评估人类与算法协作在日常工作实践中的潜力。ACM人机交互会议论文集3，计算机支持的协同工作（2019年），1–23。[158] Christine T Wolf。2019年。可解释性场景：迈向基于场景的可解释人工智能设计。载于第24届智能用户界面国际会议论文集。252–257。[159] Yao Xie、Melody Chen、David Kao、Ge Gao 和 Xiang'Anthony' Chen。2020年。CheXplain：帮助医生探索和理解数据驱动、人工智能赋能的医学影像分析。载于《2020年计算机系统中的人为因素会议论文集》。[160] Fumeng Yang、Zhuanyi Huang、Jean Scholtz 和 Dustin L Arendt。2020年。视觉解释如何帮助终端用户对机器学习建立适当信任？载于智能用户界面国际会议论文集。[161] Qian Yang。2019年。将人工智能作为用户体验设计素材进行剖析。博士学位论文。[162] Qian Yang、Aaron Steinfeld、Carolyn Rosé 和 John Zimmerman。2020年。重新审视人机交互是否、为何以及如何在设计上具有独特难度。载于计算机与人类交互会议论文集。[163] Qian Yang、Aaron Steinfeld 和 John Zimmerman。2019年。平淡无奇的人工智能：将智能决策支持融入关键的临床决策过程。载于计算机与人类交互会议论文集。[164] Qian Yang、John Zimmerman、Aaron Steinfeld、Lisa Carey 和 James F Antaki。2016年。探究心脏泵植入的决策过程：决策支持工具的助益机会。载于计算机与人类交互会议论文集。[165] Dong Whi Yoo、Michael L Birnbaum、Anna R Van Meter、Asra F Ali、Elizabeth Arenare、Gregory D Abowd 和 Munmun De Choudhury。2020年。设计面向临床医生的工具，以利用患者社交媒体活动中的洞见：迭代式共同设计方法。《JMIR心理健康》（2020年）。[166] Dong Whi Yoo、Sindhu Kiranmai Ernala、Bahador Saket、Domino Weir、Elizabeth Arenare、Asra F Ali、Anna R Van Meter、Michael L Birnbaum、Gregory D Abowd 和 Munmun De Choudhury。2021年。临床医生对利用患者社交媒体活动中的计算精神健康洞见的看法：原型的设计与定性评估。《JMIR心理健康》8卷11期（2021年），e25455。[167] Jason Yosinski、Jeff Clune、Anh Nguyen、Thomas Fuchs 和 Hod Lipson。2015年。通过深度可视化理解神经网络。arXiv预印本 arXiv:1506.06579（2015年）。[168] Quanshi Zhang、Yu Yang、Haotian Ma 和 Ying Nian Wu。2019年。通过决策树解释卷积神经网络。载于《IEEE计算机视觉与模式识别会议论文集》。6261–6270。

[169] Yunfeng Zhang、Q Vera Liao 和 Rachel KE Bellamy。2020年。置信度与解释对人工智能辅助决策中的准确性和信任校准的影响。载于公平、问责与透明度会议论文集。ACM。[170] Haiyi Zhu、Bowen Yu、Aaron Halfaker 和 Loren Terveen。2018年。价值敏感算法设计：方法、案例研究与经验教训。*ACM*人机交互会议论文集 2，计算机支持的协同工作（2018年），1–23。