A 系统的以用户为中心的可解释人工智能方法分类体系：综述 y

Ehsan Emamirad、Pouya Ghiasnezhad Omran、Armin Haller、Shirley Gregor
澳大利亚国立大学 {ehsan.emamirad, p.g.omran, armin.haller, shirley.gregor}@anu.edu.au

摘要

近年来，人工智能取得的进展与研究界不断加大的努力相伴而行，研究界致力于调查、分类和评估旨在使人工智能模型具备可解释性的各种方法。然而，现有的大多数尝试都采用以方法为中心的可解释人工智能（XAI）视角，这种视角通常仅对领域专家具有意义。目前显然缺乏一种稳健的定性与定量性能框架，用于评估解释对不同类型用户的适用性。我们综述了相关工作，并基于一般系统论的原则，提出了一种统一、包容且以用户为中心的可解释人工智能分类体系。该分类体系为我们评估可解释人工智能方法对所有用户类型（包括开发者和最终用户）的适切性提供了基础。

1 引言

近年来，为应对先进机器学习（ML）解决方案的黑箱特性，可解释人工智能方法的发展迅速，学术界也投入了大量精力研究、分类和评估可解释人工智能方法。尽管现有研究显著推动了可解释人工智能领域不断增长的研究成果，但往往对可解释人工智能概念采取割裂的视角，且受特定领域和受众的影响。此外，现有的大多数工作未必能满足新手用户的需求，因为这些工作通常直觉上以领域专家为服务对象 [Srinivasan and Chander, 2020]。此外，似乎仍缺乏健全的评估框架和指南，无法从定性和定量两方面衡量面向不同类型用户的解释是否恰当[Doshi-Velez and Kim, 2017]。先前研究中的社会科学和认知科学见解表明，只有对用户满意度、信任评估及其他社会因素进行评估，才能证明可解释人工智能的合理性 [Miller, 2019] 。

为应对上述挑战，本文提出了可解释人工智能的一种新定义，以及一种以用户为中心、基于一般系统理论原则的分类体系，旨在提供：1）一个全景式框架，使可解释人工智能更贴近人类理解；以及一个定性框架，作为基准，用于评估可解释人工智能方法对不同类型用户的适用性。

本文其余部分安排如下：第 2 节介绍背景，第 3 节梳理相关工作。第 4 节讨论可解释人工智能的系统视角分类法；第 5 节则通过评估若干可解释人工智能方法，展示该分类体系的实际应用。本文在第 6 节作结。

2 背景

尽管可解释人工智能（XAI）的理念近期受到广泛关注[Holzinger 等人，2018], 但对于可解释性的定义，学界仍未达成普遍共识，这仍是可解释人工智能研究面临的最大挑战之一。当代定义对可解释性及其应包含哪些属性，仅提供了部分且有时含糊不清的观点[Ciatto 等人， 2020]。这可能是因为，文献中的每篇综述论文都会根据自身特定的目的、领域和受众，提出各自对可解释性的定义[Vilone 和 Longo， 2020]。尽管现有定义数量庞大，但它们似乎只共享少数几项通常与科学哲学中的解释相关联的属性[Pa´ez，2020年；Ciatto 等人，2020]。例如，[Biran 和 Cotton， 2017],将可解释人工智能定义为系统解释决策和预测的能力，以及说明决策过程的透明性。另一些研究则将可解释人工智能视为一种机制，用以描述人工智能代理的优缺点，并传达对其未来行为方式的理解[Sanneman 和 Shah，2020]。

此外，现有的可解释人工智能方法大多仅限于对模型处理过程进行技术性阐述 [Lyons 等人， 2017], 这可能仍然只能满足领域专家对可解释性的需求。尽管可解释性、技术透明性和可解释性各有不同含义 [Clinciu 和 Hastie,2019]，但人们常常将它们混为一谈。可以说，可解释性关乎用户能够理解的建议所依据的理由和论证；而透明性则指理解系统中所进行的计算所关联的语义 [Pal, 2020]。同样，可解释性与解释不同：前者被视为关于模型分析基础的形式化逻辑信息，其解释性特征仅对领域专家有用[Hoffman 等人，2018]。因此，我们提出，

可解释人工智能需要重新界定，同时涵盖可解释性，透明性和可解释性，并确保所有类型的用户都能理解解释—包括领域专家和最终用户。

3 相关工作

为了概述该领域现有的综述并对其进行适当比较，我们根据这些研究所涵盖的可解释性不同方面，将先前的研究归为五类（见表1）：

1. 可解释人工智能方法。与研究和分类各种可解释性方法相关的综述，例如解释是否是内在的，或者该方法是否涉及独特的解释器（如代理分类器）；

2. 机器学习类型。侧重于机器学习要素和问题的框架与架构，包括所采用的机器学习方法类型（如神经网络）、机器学习问题（如分类）、数据类型（如非结构化数据），以及机器学习问题关联的是特征（模型前）、过程（模型内）还是输出（模型后）；3. 解释。关于解释本身的指南，包括其含义（如上下文）、格式（如文本、基于图像）和类型（如反事实、基于示例、因果、论证等）；4. 用户领域。将用户的领域知识纳入解释的开发与检索，并涵盖问题发生的背景、领域和情境的相关模型；在这些场景中，解释往往是强制性要求（如医疗领域）；5. 评估。提出并采用评估框架来衡量机器学习和可解释人工智能输出性能的研究。这些研究会区分原始模型在生成解释前后的性能。

尽管这里介绍的所有综述、框架和指南都为迅速发展的可解释人工智能研究领域作出了贡献，但据我们所知，现有综述通常提供以方法为中心的框架，用于分类和评估可解释人工智能模型——也就是说，围绕可解释人工智能这一概念的各种概念和要素，通常都是依据方法本身进行考察和分类。此外，除两项研究（如表1所示）强调了最终用户的需求和要求、他们的领域知识与背景，以及他们与人工智能代理交互以生成解释时所处的情境之外，没有其他综述从独特视角对用户可理解的可解释性进行分类。这与此前研究的发现一致：这些研究证实，在构建可解释人工智能应用时，缺少人在回路方面的要求 [Chander 等人，2018]。

4 **系统**的可解释人工智能视角

为弥补上述空白，本综述借鉴信息系统（IS）研究界，尤其是一般系统理论的核心概念 [Bertalanffy 等人，1968]，提出了一种新颖且独特的可解释人工智能视角。本研究有两个目标：第一，提出一个以用户为中心的可解释人工智能分类体系及其环境，在其中，用户与人工智能代理互动时，不仅是解释的接收者，也是生成解释的关键参与者。第二，构建一个统一、整体且包容的框架，为评估可解释人工智能方法对各类用户的适用性提供依据。为验证所提分类体系的可行性，我们对广泛的可解释人工智能方法进行了全面的元分析，并对综述中发现的、影响可解释人工智能方法设计的一系列相互关联要素进行了分类。这些要素包括解释的概念、可解释性的类型、可解释性的属性、应用领域、人在回路、可解释人工智能方法与技术，以及性能与评估。

将这些要素结合起来，我们可以依据它们在设计可解释人工智能架构时的相互关系及其潜在作用，对其进行概念化。图1通过相应的概念性元模型，展示了各种可解释人工智能要素之间的相互依赖关系。下文将以给定的问题领域为例：一种典型的高性能黑箱机器学习方法（如深度神经网络 [Gunning and Aha，2019]，）需要一种解释，而这种解释通常通过特定的可解释人工智能方法或技术生成。所需的解释应包含用户易于理解的关键类型和属性。最终用户还会评估原始机器学习模型和可解释人工智能模型的性能，以判断解释是否易于理解。用户的类型和背景知识、需要解释的情境、用户在生成解释或理解解释时与人工智能代理的参与和交互，以及生成这些解释所付出的努力，也都会在判断可解释人工智能的结果是否合适、进而是否值得信任方面发挥关键作用 [Gregor and Yu，2002]。研究人员指出，信任也是人工智能应用获得广泛接受的必要条件，尤其是在医疗领域等要求预测结果具有充分论证和可靠性的场景中 [Holzinger 等人，2018]。

应用一般系统论的原则，我们可以进一步推断，元模型中的概念是一组相互关联、彼此联系的要素，这些要素之间具有足够的连贯性和交互性，使其作为一个整体具备实用性和意义。在此，整体性视角要求我们将可解释人工智能视为处于交互中的部分系统，而不是一个独立存在的概念、方法或技术。因此，本研究进一步提出如下对可解释人工智能的新定义：

定义1：给定特定应用领域，可解释人工智能是一个由多个要素构成的系统，并配备相应的方法和行为，以确保人工智能代理（包括其生成的输出）能够被相关用户理解并获得信任。

除了为可解释人工智能提出一种全新的视角之外，将其概念纳入系统视角分类法，还将在诸多方面展现出独特性。首先，现有的可解释人工智能方法与技术可以基于输入—过程—输出（IPO）模型进行进一步分类，而该模型已被广泛

研究/类别	可解释性技术				机器学习类型				解释					User 领域		评估
[Adadi and Berrada，2018]	✓	✓	✓													
[Anjomshoae 等，2019]				✓						✓	✓	✓			✓	✓
[Arrieta 等，2020]	✓		✓		✓				✓			✓				
[Carvalho 等，2019]	✓	✓	✓	✓		✓			✓	✓		✓				
[Chakraborti 等，2020]		✓							✓							✓
[Das 和 Rad，2020]	✓	✓			✓		✓	✓								✓
[Dos˘ilovic´ 等，2018]	✓															
[Du 等，2020]	✓	✓														
[Gilpin 等，2018]				✓					✓		✓	✓				
[Guidotti 等，2019b]	✓	✓	✓	✓	✓		✓	✓	✓			✓				
[Li 等，2022]	✓				✓			✓				✓				
[Mohseni 等，2021]	✓	✓	✓						✓			✓		✓		✓
[Nunes 和 Jannach，2017]				✓	✓		✓	✓	✓	✓			✓	✓	✓	
[Zhang 和 Chen，2020]				✓	✓		✓	✓				✓				✓
[Vilone 和 Longo，2020]	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓			✓

表1：考察的相关综述。

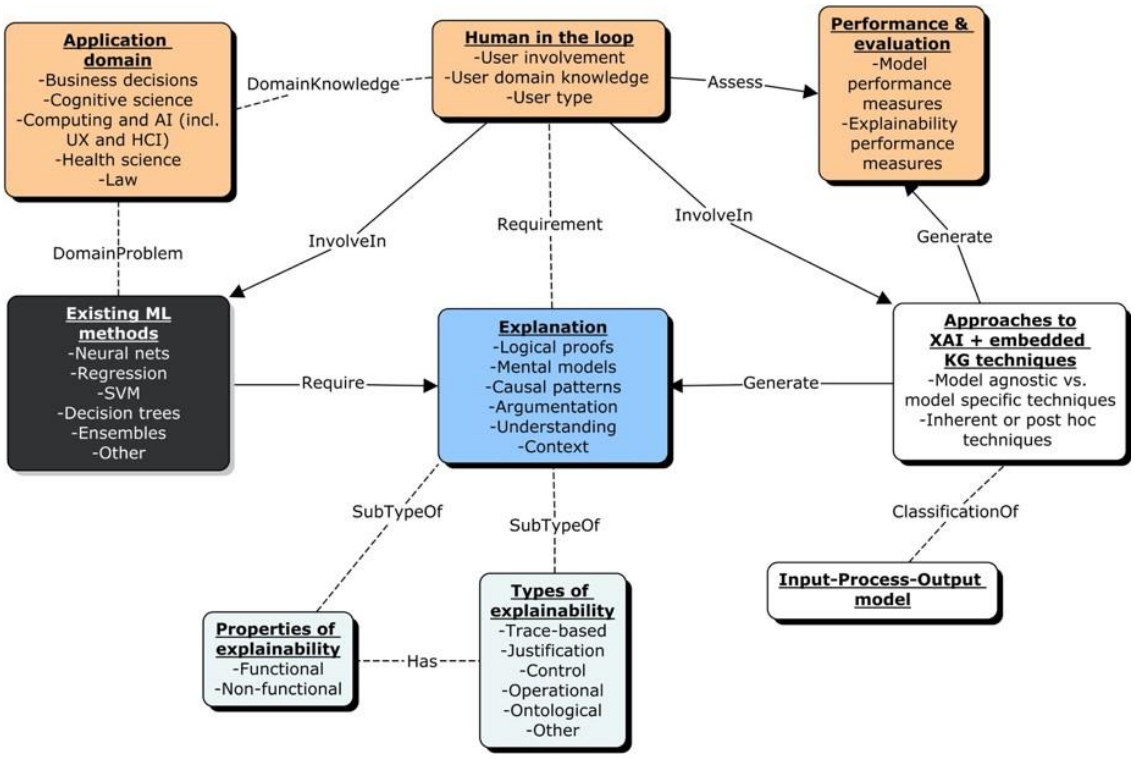

图1：提出的可解释人工智能元模型。

信息系统理论和实践中采用的方法。此处相关的机会在于，探究能否让机器学习方法在输入层面（模型前）、过程层面（模型内）或输出层面（模型后）变得可解释。这种分类方式类似于理解模型是否本质上可解释（过程层面的可解释性），还是需要使用事后可解释性技术（输出层面的可解释性）。其次，这一视角提供了一种以用户为中心的方法，让各类用户无论具备何种背景知识，都能在人工智能代理的开发、理解和建立信任的整个过程中参与其中并进行互动。第三，解释的本质及其类型和属性共同构成了可解释性的概念。此外，在判断解释是否恰当时，可解释性性能被视为与原始模型的性能同等重要。最后，可以建立定性基准，以帮助判断某种可解释人工智能方法是否易于用户理解，或识别其不足之处，从而加以改进，使该方法易于用户理解。下文将简要回顾上述各个要素。

4.1 解释解释理解并界定解释

是一项理论与实践上的挑战，因为解释具有情境性，且在不同领域中含义各异 [Craik，1952]。对解释的研究在哲学领域有着悠久的传统，且主要与认知理论和直觉理论相关；它也为人工智能带来了重要的新视角 [Fagin等人，2004], 尤其是在探究人们期待系统给出的解释是否与人们通常期待其他人给出的解释相符方面 [Gregor and Yu，2002]。

关于解释的文献浩如烟海，内容也极为多样，大量研究致力于演绎证明、因果关系模式、解释性立场与心理立场、论证以及理解。在当代关于解释的研究中，一组基本法律被作为公理提出，而对这些法律的演绎证明则被视为解释 [Keil，2006]。解释也被视为追溯探究因果关系时的推理过程。在此，解释是源自特定事件中事实的逻辑事件 [Von Wright，1971]。此外，解释也被看作心智模型，其形式可以从逻辑模式的形式化表征到系统的类图像内部表征不等 [Norman等人，1983]。无论解释的本质如何，它们都通过构成论证、概括和建构解释的基础来促进理解 [Wiggins等人，2005]。

4.2 可解释性的类型与解释本身一样，许多

近期文献大多致力于讨论解释的分类。许多研究提出，解释要被视为全面，就应回答若干问题 [Vilone 和 Longo，2020], ，主要包括基于追踪的问题 [Gregor 和 Yu，2002年；Preece，2018]。解释类型是否适用，还取决于其内容、格式、预期生成时间，以及用户在决定解释类型时的参与程度。

文献中提及的大多数解释类型，似乎都局限于最常见的“如何”和“为什么”这类查询 [Mohseni 等，2021]。例如，对比式解释和反事实解释旨在为所生成的解释提供论证、论据或反驳论据[Hoffman 等人，2018]。为回答“为什么”这类查询，研究者还会提供机制性解释和因果解释；从本质上说，这些解释与众多研究中讨论的基于追踪的解释相关 [Abdul 等，2018]。此外，还有基于案例、情境和示例的解释，它们与解释的语境以及相似案例或示例相关，有助于阐明解释中的“如何”这一方面 [Cai 等，2019]。

4.3 可解释性的属性可解释性的概念也处于

若干属性的交汇处，这些属性往往取决于机器学习模型和用户的特征（[Doran等人，2017]）。回顾相关文献可见，其中一些属性可视为功能性属性，例如保真度，它衡量解释对黑箱行为的近似是否可靠、准确（[Pedreschi等人，2019], ）；另一些则被视为非功能性属性，包括透明性、理解、推理和可信度等（[Nunes 和 Jannach，2017]）。有人认为，无论哪种情况，人工智能代理若不同时支持这两类属性，就不具备可解释性（[Lipton，2018年；Hoffman 等人，2018年；Hagras，2018年；Doran等人，2017]）。

4.4 应用领域

生成解释时，需要仔细考虑应用领域及其情境 [Rudin, 2019], , 并据此考量用户的特性以及用户在生成解释时所扮演的角色。解释的生成应基于解释的复杂程度与人类能力和领域知识的复杂程度之间的匹配关系 [Collaris 和 van Wijk, 2020]。此外，解释还应根据与人工智能代理互动的人所扮演的角色，以及用户参与解释生成的程度进行调整 [Miller, 2019]。

文献中详细探讨了众多应用领域。其中包括认知科学领域的解释，这类解释通常被描述为一种交互式知识迁移过程 [Srinivasan 和 Chan-der, 2020；Miller，2019；Floridi 等人，2020；Gregor and Yu, 2002]；计算机科学领域的解释，关注模型的分析性和数学透明性 [Lyons 等人，2017；Hoffman 等人，2018]；人机交互领域的解释，通常与用户对系统的感知和理解力相关 [Sanneman 和 Shah，2020；Pa´ez, 2020；Miller，2019；Hagras, 2018]；以及健康科学和法律等影响重大的领域。在这些领域中，可信度、因果性、信息丰富度、公平性、可访问性和隐私意识被认为是关键因素 [Biran 和 Cotton，2017；Arrieta 等人，2020；Felzmann 等人，2019]。

4.5 人在回路

机器学习模型的可解释性在很大程度上取决于其预期用户、用户的类型和背景知识、用户参与机器学习模型训练的程度，以及

处理解释的过程，以及他们为获取解释所付出的努力，还有他们基于自身认知行为能力、局限性和领域知识理解输出的能力[Gregor and Yu，2002 年；Ribera and Lapedriza，2019]。开发用户可理解的可解释人工智能，需要了解寻求解释的人类参与者的需求，并让他们尽早参与可解释模型的设计过程——而现有可解释人工智能工作明显缺少这一关键条件[Srinivasan and Chander，2020]。

就用户类型而言，针对不同用户制定了多项可解释性要求；在设计可解释模型时，应对这些要求加以考虑并予以区分。例如，对于领域专家或人工智能开发者而言，可解释性要求可以描述为理解系统是否正常运行，以便识别并消除错误；而对于最终用户或接收这些解释的新手用户而言，同样的要求则相反地表现为理解系统行为的推理和论证，从而建立信任[Felzmann等人，2019]。此外，最终用户还负责通过处理、整理和理解人工智能代理的输出，对其所推荐的决策作出判断；这一过程应在很大程度上与其认知上的“直觉”相对应[Doran等人，2017]。为了使代理的行动易于理解，还必须考虑最终用户的特定情境、背景知识、兴趣和局限性[Hagras，2018]——最终用户应尽可能充分地理解代理及其输出的预测可靠性[Pa´ez，2020]。

4.6 可解释人工智能方法与技术

旨在有效实现可解释人工智能的方法日益增多，参见例如[Guidotti 等，2019b；Gilpin 等，2018年；Miller，2019；Adadi and Berrada，2018年；Vilone 和 Longo，2020年；Zhang等人，2022]。现有的可解释人工智能方法可分为两大类：一类是本质上可解释、在设计上便于解释的方法，例如符号人工智能、决策树、回归等；另一类是事后可解释性方法，旨在从不透明的黑箱模型中提取易于理解的信息，例如代理模型、特征提取方法、深度解释等。

现有方法还可以根据其提供的是与单个实例相关的解释，还是针对整个预测模型的解释，进一步分类。局部实例解释（也称为特定于模型的方法）仅针对单个预测提供局部近似。而全局模型解释（即模型无关方法）则提供黑箱模型的近似，旨在实现对整个模型的全局复现和理解。最后，可解释人工智能方法还可根据用户参与生成解释的程度进行分类。静态方法只是向最终用户呈现解释，而交互式方法则让用户有机会通过视觉内容、对话、规则等输出格式与解释进行互动。

4.7 性能与评估

可解释人工智能的主要目标之一是提升人工智能代理的性能、学习能力和用户感知[Gregor and Yu, 2002]。然而，代理的可解释性与性能之间往往需要权衡。尽管文献中已有关于这些权衡的讨论，例如参见[Rudin, 2019],但普遍认为，随着人工智能代理的内部复杂性增加，模型似乎会变得更难解释[Ha- gras, 2018]。因此，接下来的任务是设计和开发可解释人工智能方法，在保持较高预测准确率的同时实现可解释性[Holzinger 等人，2017]，，并在信息完整性与用户可理解性之间取得适当平衡[Pa´ez, 2020]。另一项挑战是确定解释是否有效，以及最终用户是否已对人工智能代理形成足够的理解和信任。在这种情况下，重要的是设计更具体且覆盖面更广的定性衡量指标，以评估解释的质量、解释满意度、对系统的信任，以及其他适当的社会因素[Hoffman 等人，2018年; Miller, 2019]。

5 可解释人工智能方法的评估

如前所述，现有框架未必是为评估可解释人工智能方法的性能而设计的，尤其是在用户理解度方面。受上述讨论启发，本文展示所提出的分类体系在文献中一系列可解释人工智能方法上的实际应用。表 2 总结了这些方法；选择它们时考虑了若干因素：1）鉴于用户相关考量，将分类体系应用于多种方法和问题领域；2）回顾并分类兼具广泛适用性的方法（如代理分类器）和适用范围较窄的方法（如深度张量）；3）这些方法在文献中的可信度和相关性。对这些方法的评估带来了若干重要启示：

用户考量。所有方法似乎都要求用户具备与领域专家相当的背景知识。因此，这些方法生成的解释是否能具备易于新手用户理解的特性，尚不明确。大多数可解释人工智能方法的应用领域是计算，这进一步印证了这一点。另一个值得注意的观察是，除基于知识的方法和 LIME 外，现有的大多数研究都将领域专家视为目标用户类型。这与要求特定用户（例如领域专家）参与生成解释的做法一致，但并未明确说明这些解释是否满足新手用户的需求。

解释。大多数受评方法与模型无关，旨在为单个预测提供解释。同样，大多数方法通过事后可解释性技术（例如部署代理分类器）在输出层级提供解释，而少数方法则侧重于通过特征重要性提取等技术，揭示机器学习模型的内部运作，从而提高其透明性。在这方面，仍有机会进一步探究

选定的可解释人工智能方法		分类体系						非功能性				其他			
		分类	模型	范围	局部性	解释	协同过滤	性能	保真度	可理解性	鲁棒性	用户	专业知识	可视化	上下文
[Angelov 和 Soares，2020]	*xDNN*	PR	M	Agn	Loc	R		✓	✓			H	E		
[Anjomshoae 等，2019]	CI&CU	MC	B	Agn	Loc	J, C		✓	✓	✓		H	E	✓	
[Casalicchio 等人，2018]	SFIMP	FI	C	Agn	Both	R			✓	✓		H	E	✓	
[Shrikumar 等人，2017]	DeepLIFT	FI	C	Agn	Loc	R			✓			H	E	✓	✓
[Guidotti 等，2019a]	*LORE*	SM	C	Agn	Loc	R, C	✓		✓	✓		M	E	✓	
[Collaris 和 van Wijk，2020]	XExplore	SM	C	Agn	Both	R	✓		✓	✓		H	E	✓	✓
[Ribeiro 等人，2016]	*LIME*	SM	C	Agn	Loc	R	✓		✓		✓	M	N	✓	✓
[Ehsan 等人，2019]	*ARM*	TS	C	Agn	Both	R, J		✓	✓	✓		M	E		
[Lou 等人，2013]	*GA2M*	–	C	Sp	Loc	R			✓			H	E	✓	
[Ai 等人，2018]	–	KB	C	Agn	Both	R	✓		✓	✓		L	N		✓
[Fuji 等人，2019]	深度张量	KB	M	Agn	Both	R, J	✓		✓	✓	✓	M	N	✓	✓

表 2：选定的可解释人工智能方法。PR = 基于原型，MC = 多决策准则，FI = 特征重要性，SM = 代理模型，TS = 训练解释集，KB = 基于知识 | B = 商业应用，M = 医学，C = 计算机科学 | NN = 《神经网络》，LRM = 线性回归模型，SVM = 支持向量机，RFC = 随机森林分类器，GLM = 一般线性模型，NB = 朴素贝叶斯，DT = 决策树，ACLF = 所有分类器 CF = 协同过滤 | Agn = 不可知，Sp = 特定，Loc = 局部 | R = 推理，J = 论证，C = 对比式 | H = 高，M = 中，L = 低 | E = 专家，N = 新手

将输入级可解释性进一步融入可解释人工智能方法。在所调查的所有方法中，识别并解释预测决策的因果模式，以及阐明预测行为的推理过程，似乎是两种主要的解释类型。就解释的属性而言，虽然只有少数方法将保真度和忠实度视为必要的功能属性，但在所有方法中，理解被证明是最重要的非功能属性。

性能。一些方法提供了度量指标或一组评估指标，用于展示解释生成前后相较于原始模型的性能提升，具体包括更高的预测准确率、更短的训练时间、更低的计算资源消耗、更小的结果差异，以及更高的精确率和召回率。然而，没有一种方法讨论模型性能与可解释性性能之间的权衡，尤其是在评估解释是否恰当且易于最终用户理解时。与前文相同，基于知识的方法和 LIME 方法是仅有的能够提供性能度量指标的方法，这些指标在一定程度上满足最终用户对可解释性的要求。

6 结论

展望未来，亟需制定评估策略，量化用户在解释提供前后与可解释人工智能方法互动时的行为。其目的不仅是从定性和定量两方面衡量解释的性能，还要从最终用户的角度，确定用户是否具备实用的——理解能力并信任人工智能代理。此外，构建与基于知识的系统集成、独立于领域的可解释人工智能架构，将是一种值得进一步研究的重要范式。其目标是将基于知识的系统嵌入现有人工智能模型，从而为不同类型的用户提供更易于理解的解释。

本文综述了一系列可解释人工智能研究与方法，并对围绕可解释人工智能这一概念的各种要素和概念进行了详细的元综述。我们还基于系统理论以及全视角、以用户为中心的可解释人工智能分类体系，提出了一种全新的可解释人工智能视角。最后，我们将该分类体系应用于文献中选取的几种可解释人工智能方法，通过定性评估其对用户的适用性，证明了该分类体系的可行性，而不受用户领域知识的限制。

参考文献[

Abdul 等人，2018] Ashraf M. Abdul、Jo Ver meulen、Dand- ing Wang、Brian Y. Lim 和 Mohan S. Kankanhalli。《面向可解释、可问责和可理解系统的趋势与发展轨迹：HCI 研究议程》。载于 CHI，第 582 页，2018年。[Adadi and Berrada，2018] Amina Adadi 和 Mohammed Berrada。《窥探黑箱内部：可解释人工智能（XAI）综述》。IEEE Access，6:52138–52160，2018年。[Ai 等人，2018] Qingyao Ai、Vahid Azizi、Xu Chen 和 Yongfeng Zhang。《学习异构知识库嵌入以实现可解释推荐》。算法，11(9):137，2018年。

[Angelov, Soares, 2020年] Plamen P Angelov 和 Eduardo A. Soares。迈向可解释的深度神经网络（xdnn）。《神经网络》，130:185–194，2020年。

[Anjomshoae 等， 2019] Sule Anjomshoae、Kary Fra¨mling 和 Amro Najjar。通过情境重要性和效用解释黑箱模型的预测。载于 EXTRAAMAS，Lecture Notes in Computer Science 第 11763 卷，第 95–109 页，2019年。[Arrieta 等，2020] Alejandro Barredo Arrieta、Na- talia D´ıaz Rodr´ıguez、Javier Del Ser、Adrien Bennetot、Siham Tabik、Alberto Barbado、Salvador Garc´ıa、Sergio Gil-Lopez、Daniel Molina、Richard Benjamins、Raja Chatila 和 Francisco Herrera。可解释人工智能（XAI）：迈向负责任人工智能的概念、分类体系、机遇与挑战。《信息融合》，58:82–115，2020年。[Bertalanffy 等人，1968] Anatol Rapoport Bertalanffy、Ken- neth E Boulding、William Ross Ashby、Margaret Mead 和 Gregory Bateson。L. von Bertalanffy，《一般系统论》。纽约：George Braziller，1968年。[Biran 和 Cotton，2017] Or Biran 和 Courtenay Cotton。机器学习中的解释与论证：一篇综述。载于 IJCAI-17 可解释人工智能（XAI）研讨会，第 8 卷，第 8–13 页，2017年。[Cai 等人，2019] Carrie J. Cai、Jonas Jongejan 和 Jess Hol- brook。基于示例的解释对机器学习界面的影响。载于 IUI，第 258–262 页，2019年。[Carvalho 等， 2019] Diogo V Carvalho、Eduardo M Pereira 和 Jaime S Cardoso。机器学习的可解释性：方法与度量指标综述。《电子学》，8(8):832，2019年。[Casalicchio 等人，2018] Giuseppe Casalicchio、Christoph Molnar 和 Bernd Bischl。黑箱模型特征重要性的可视化。载于 ECML PKDD，第 655–670 页，2018年。[Chakraborti 等， 2020] Tathagata Chakraborti、Sarath Sreedharan 和 Subbarao Kambhampati。可解释自动规划与决策领域的新兴格局。载于 IJCAI，第 4803–4811 页。ijcai.org，2020年。[Chander 等人， 2018] Ajay Chander、Ramya Srinivasan、Suhas Chelian、Jun Wang 和 Kanji Uchino。与信念共事：企业中的人工智能透明度。载于 IUI Workshops，2018年。[Ciatto 等人， 2020] Giovanni Ciatto、Michael Ignaz Schumacher、Andrea Omicini 和 Davide Calvaresi。人工智能中的基于智能体的解释：迈向抽象框架。载于 EXTRAAMAS，Lecture Notes in Com- puter Science 第 12175 卷，第 3–20 页，2020年。[Clinciu 和 Hastie，2019] Miruna-Adriana Clinciu 和 He- len Hastie。可解释人工智能术语综述。载于 NL4XAI，第 8–13 页，2019年。[Collaris 和 van Wijk，2020] Dennis Collaris 和 Jarke J. van Wijk。Explainexplore：机器学习解释的视觉探索。载于 IEEE，第 26–35 页，2020年。

[Craik，1952] Kenneth James Williams Craik。《解释的本质》，第 445 卷。CUP Archive，1952年。[Das 和 Rad，2020] Arun Das 和 Paul Rad。机遇以及可解释人工智能（可解释人工智能）领域的挑战：一篇综述。arXiv 预印本 arXiv:2006.11371，2020年。

[Doran等人，2017] Derek Doran、Sarah Schulz 和 Tarek R Besold。可解释人工智能究竟意味着什么？一种新的视角概念化。arXiv 预印本 arXiv:1710.00794，2017年。

[Doshi-Velez 和 Kim，2017] Finale Doshi-Velez 和 Been Kim。迈向可解释机器学习的严谨科学。arXive 预印本 arXive:1702.08608，2017年。

[Dos˘ilovic´, Dos˘ilovic´, Mario Brc˘ic´ 和 Nikica Hlupic´。可解释人工智能：一篇综述。载于 MIPRO，第 0210–0215 页，2018年。等，2018] Filip Karlo

[Du 等， 2020] Mengnan Du、Ninghao Liu 和 Xia Hu。可解释机器学习技术。Commun. ACM，63(1):68–77，2020年。[Ehsan 等人，2019] Upol Ehsan、Pradyumna Tambwekar、Larry Chan、Brent Harrison 和 Mark O. Riedl。自动生成理由：一种可解释人工智能技术及其对人类感知的影响。载于 IUI，第 263–274 页，2019年。[Fagin等人，2004] Ronald Fagin、Joseph Y Halpern、Yoram Moses 和 Moshe Vardi。关于知识的推理。麻省理工学院出版社，2004年。[Felzmann 等人，2019] Heike Felzmann、Eduard Fosch Vil- laronga、Christoph Lutz 和 Aurelia Tamo`-Larrieux。值得信赖的透明性：人工智能在法律规范与情境相关问题之间的透明性要求。《大数据与社会》，6(1):2053951719860542，2019年。[Floridi 等人， 2020] Luciano Floridi、Josh Cowls、Thomas C. King 和 Mariarosaria Taddeo。如何为社会公益设计人工智能：七个关键因素。《科学与工程伦理》，26(3):1771–1796，2020年。[Fuji 等人， 2019] Masaru Fuji、Hajime Morita、Keisuke Goto、Koji Maruhashi、Hirokazu Anai 和 Nobuyuki Igata。通过深度张量与知识图谱的结合实现可解释人工智能。Fujitsu Sci. Tech. J，55(2):58–64，2019年。[Gilpin 等， 2018] Leilani H. Gilpin、David Bau、Ben Z. Yuan、Ayesha Bajwa、Michael A. Specter 和 Lalana Kagal。解释解释：机器学习可解释性概述。载于 IEEE，第 80–89 页，2018年。[Gregor and Yu，2002] Shirley Gregor 和 Xinghuo Yu。探索智能系统技术的解释能力。载于模糊逻辑，第 288–300 页。2002年。[Guidotti 等，2019a] Riccardo Guidotti、Anna Monreale、Fosca Giannotti、Dino Pedreschi、Salvatore Ruggieri 和 Franco Turini。黑箱决策的事实性与反事实解释。IEEE Intell. Syst.，34(6):14–23，2019年。

[Guidotti 等，2019b] Riccardo Guidotti、Anna Monreale、Salvatore Ruggieri、Franco Turini、Fosca Giannotti 和 Dino Pedreschi。解释黑箱模型的方法综述。ACM Comput. Surv.，51(5):93:1–93:42，2019年。[Gunning and Aha，2019] David Gunning 和 David W. Aha。DARPA 的可解释人工智能（XAI）计划。AI Mag.，40(2):44–58，2019年。[Hagras，2018] Hani Hagras。迈向易于人类理解的可解释人工智能。Computer，51(9):28–36，2018年。[Hoffman 等人，2018] Robert R. Hoffman、Shane T. Mueller、Gary Klein 和 Jordan Litman。可解释人工智能的度量指标：挑战与前景。CoRR，abs/1812.04608，2018年。[Holzinger 等人，2017] Andreas Holzinger、Chris Biemann、Constantinos S. Pattichis 和 Douglas B. Kell。构建用于医疗领域的可解释人工智能系统需要什么？CoRR，abs/1712.09923，2017年。[Holzinger 等人，2018] Katharina Holzinger、Klaus Mak、Peter Kieseberg 和 Andreas Holzinger。我们能信任机器学习的结果吗？安全关键型决策支持中的人工智能。ERCIM News，2018(112)，2018年。[Keil，2006] Frank C Keil。解释与理解。Annu. Rev. Psychol.，57:227–254，2006年。[Li 等，2022] Xiao-Hui Li、Caleb Chen Cao、Yuhan Shi、Wei Bai、Han Gao、Luyu Qiu、Cong Wang、Yuanyuan Gao、Shenjia Zhang、Xun Xue 和 Lei Chen。数据驱动与知识感知的可解释人工智能综述。IEEE Trans. Knowl. Data Eng.，34(1):29–49，2022年。[Lipton，2018] Zachary C. Lipton。模型可解释性的迷思。ACM Queue，16(3):30，2018年。[Lou 等人，2013] Yin Lou、Rich Caruana、Johannes Gehrke 和 Giles Hooker。具有成对交互作用的准确、可理解模型。载于 ACM SIGKDD，第 623–631 页，2013年。[Lyons 等人，2017] Joseph B. Lyons、Matthew A. Clark、Alan R. Wagner 和 Matthew J. Schuelke。自主系统中的可认证信任：让棘手问题变得具体可感。AI Mag.，38(3):37–49，2017年。[Miller，2019] Tim Miller。人工智能中的解释：来自社会科学的见解。Artif. Intell.，267:1–38，2019年。[Mohseni 等，2021] Sina Mohseni、Niloofar Zarei 和 Eric D Ragan。可解释人工智能系统设计与评估的多学科综述及框架。ACM Transactions on Interactive Intelligent Systems (TiiS)，11(3-4):1–45，2021年。[Norman等人，1983] Donald Norman、Dedra Gentner 和 AL Stevens。心智模型。Human-computer Interaction，章节，第 7–14 页，1983年。[Nunes 和 Jannach，2017] Ingrid Nunes 和 Dietmar Jannach。决策支持系统与推荐系统中的解释：系统综述与分类体系。User Model. User Adapt. Interact.，27(3-5):393–444，2017年。[Pa´ez, 2020] Andre´s Pa´ez. 可解释人工智能（XAI）中的实用主义转向。CoRR, abs/2002.09595, 2020.[Pal, 2020] Nikhil R. Pal. 探寻具备类人认知与推理能力、值得信赖的透明智能系统。Frontiers Robotics AI, 7:76, 2020.[Pedreschi 等人，2019] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. 黑箱人工智能决策系统的有意义解释。载于 AAAI, 第 9780–9784 页, 2019.[Preece, 2018] Alun D. Preece. 在人工智能中追问“为什么”：智能系统的可解释性——观点与挑战。Intell. Syst. Account. Finance Manag.，25(2):63–72, 2018.[Ribeiro 等人, 2016] Marco Tu´lio Ribeiro, Sameer Singh, and Carlos Guestrin. “我为什么应该相信你？”：解释任意分类器的预测。载于 ACM SIGKDD, 第 1135–1144 页, 2016.[Ribera and Lapedriza，2019] Mireia Ribera and Agata Lapedriza. 我们能否提供更好的解释？以用户为中心的可解释人工智能提案。载于 IUI Workshops, 第 2327 卷, 第 38 页, 2019.[Rudin，2019] Cynthia Rudin. 对高风险决策，停止解释黑箱机器学习模型，改用可解释模型。Nat. Mach. Intell., 1(5):206–215, 2019.[Sanneman 和 Shah，2020] Lindsay M. Sanneman 和 Julie A. Shah. 基于情境意识的可解释人工智能设计与评估框架。载于 EXTRAA-MAS,《Lecture Notes in Computer Science》第 12175 卷, 第 94–110 页, 2020.[Shrikumar 等人, 2017] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 通过传播激活差异学习重要特征。载于 ICML,《Proceedings of Machine Learning Research》第 70 卷, 第 3145–3153 页, 2017.[Srinivasan and Chander, 2020] Ramya Srinivasan and Ajay Chander. 来自认知科学的解释视角——综述。载于 IJCAI, 第 4812–4818 页, 2020.[Vilone 和 Longo, 2020] Giulia Vilone 和 Luca Longo. 可解释人工智能：系统综述。arXive 预印本 arXive:2006.00093, 2020.[Von Wright, 1971] Georg Henrik Von Wright. 解释与理解。1971年.[Wiggins等人, 2005] Grant Wiggins, Grant P Wiggins, and Jay McTighe. 理解为设计。Ascd, 2005年.[Zhang 和 Chen, 2020] Yongfeng Zhang and Xu Chen. 可解释推荐：综述与新视角。Found. Trends Inf. Retr., 14(1):1–101, 2020.[Zhang等人，2022] Ziwei Zhang, Peng Cui, and Wenwu Zhu. 图深度学习：综述。IEEE Trans. Knowl. Data Eng., 34(1):249–270, 2022.