人-无人车交互中的可解释性交互研究

郭炜炜，王琦
（同济大学，上海 200092）

摘要：目的 随着现代人工智能技术在自动驾驶系统中的广泛应用，其可解释性问题日益凸显，为此探讨人-无人车交互过程中的可解释性交互的框架以及设计要素等问题，以增强自动驾驶系统的决策透明性、安全性和用户信任度。**方法** 结合可解释人工智能和人机交互的基本理论与方法，本文首先介绍了可解释性人工智能，对当前可解释内容的提取方法进行总结，然后以人-机器人交互的透明度模型为基础，建立人-无人车交互中可解释性交互的框架。最后从解释的对象、方式和评价等多个设计维度对可解释性的交互设计问题进行探讨，并结合案例进行分析。**结论** 可解释性作为人与模型决策之间的接口，不仅仅是一个人工智能技术问题，而且与人密切相关，涉及到人-无人车交互中的多个层次。本文提出人-无人车交互中可解释性交互的框架，得出在人-无人车交互每个阶段需要的解释内容以及可解释交互设计的要素。

关键词：人车共驾；人-无人车交互；可解释性人工智能；透明度
中图分类号：TB472 **文献标识码：**A **文章编号：**1001-3563(2020)18-0022-07
DOI：10.19554/j.cnki.1001-3563.2020.18.003

Explainable Interaction in Human-Autonomous Vehicle Interaction

GUO Wei-wei, WANG Qi
(Tongji University, Shanghai 200092, China)

ABSTRACT: With the wide application of modern artificial intelligence technology in autonomous driving system, its interpretability problem has become increasingly prominent. The work aims to discuss the explainability and interpretability of the autonomous systems during the interaction between the human and autonomous vehicles, in order to enhance the decision-making transparency, safety and user trust of the autonomous driving systems. Combining the fundamental theory and methodology of explainable artificial intelligence and human-computer interaction, explainable artificial intelligence (XAI) was firstly introduced and then the current methods and techniques of XAI generating the explanations were summarized. Based on the transparency model of human-robot interaction, the interaction framework for the explanations between human and autonomous vehicles was established. Finally, interaction design for explanations was discussed from the aspects of users, form and evaluation. As an interface between people and decision model, explainability is not only a domain of artificial intelligence but closely related to the people, involving multiple levels of the human-autonomous vehicle interactions (HAI). This paper proposes the interaction framework for explainability of HAI and obtains the contents requiring explanation in each stage of HAI and elements that can explain the interactive design
KEY WORDS: human-autonomous vehicle co-driving; human-autonomous vehicle interaction; explainable artificial intelligence; transparency

近年来，随着人工智能与自动驾驶技术的蓬勃发展，无人车正在逐渐走进人们的视野和生活中[1]。例如，特斯拉推出的 Autopilot 自动驾驶辅助系统可以在特定场合取代人类进行自主驾驶。尽管以深度学习

收稿日期：2020-07-29
基金项目：中央高校基本科研业务经费专项资金项目（22120190209）
作者简介：郭炜炜（1983—），男，江苏人，博士，同济大学助理教授，主要从事模式识别与图像理解和人机交互等方面的研究。
通信作者：王琦（1988—），女，河北人，博士，同济大学助理教授，主要从事可穿戴技术和人机交互方面的研究。

为代表的现代人工智能技术极大地提升了智能驾驶系统的自主感知、学习和行动的能力，但其内部过程犹如黑盒子，可解释性差，使得用户很难理解其背后的工作机理和决策逻辑，难以掌握系统决策行为的边界，人机沟通交流和协作存在较大的障碍[2]。特别是在较为复杂的行车环境下，基于机器学习与深度学习的车辆自主感知与决策往往存在较大的不确定性，存在较大的决策风险和安全隐患。可解释性旨在对于模型的工作机理和决策逻辑给出令人理解的清晰的概括和指示，使驾驶员能够及时发现自动驾驶系统可能的决策错误与漏洞而采取紧急的干预措施。这就迫切需要在人车交互过程中实现可解释的人机沟通与协作，这对于用户理解、信任和管理模型，增强自动驾驶系统安全性、决策透明度和用户信任度至关重要，对于促进无人驾驶的落地应用也具有重要的意义[3]。可解释性不仅是揭示模型决策背后的机理、逻辑以及可能的缺陷与漏洞，而且作为模型与人的接口，与用户的类型、认知能力和知识背景等方面密切相关，因而受到人机交互领域的广泛关注[4]。

本文首先从技术的视角讨论可解释性人工智能的技术和方法，然后从人机交互的视角，提出人-无人车的可解释性交互框架，并从解释的对象、方式和评价等多个设计维度进行探讨，最后对本文进行总结和展望。

1　可解释性人工智能

随着人工智能技术在诸多领域的广泛应用，其可解释性问题日益受到政府、学术界和工业界的广泛重视。例如美国国防部高级研究计划署（Defense Advanced Research Projects Agency, DARPA）启动了一项名为可解释性人工智能（Explainable Artificial Intelligence, XAI）的大型项目[5]，我国也在《新一代人工智能发展规划》中，明确将“实现具备高可解释性、强泛化能力的人工智能”作为未来我国人工智能发展的重要突破口。欧盟则通过立法规定人们对算法的决策有“要求解释的权利”。

1.1　什么是可解释性

可解释性通常被定义为向人类解释或呈现可理解的术语的能力[6]。从本质上讲，可解释性分为可理解和可判读两个方面。可理解是指机器学习模型或者智能代理内部工作机理透明，能对决策结果和行为给出相应的解释和原因，揭示其背后的决策逻辑，并能发现其中的逻辑漏洞和错误；可判读是指能够清晰地掌握机器模型或者智能代理的行为边界，结果可预测。当前的自动驾驶系统大量地采用了机器学习与深度学习技术进行环境感知、决策、规划、控制与执行。这就需要将这些“黑盒”模型转化为“白盒”模型。Koo 等人的研究表明如果自动驾驶系统能够向用户提供导致其决策行为的原因，能够显著地增强用户对自动驾驶系统的信任度[7]。自动驾驶的可解释性系统见图 1，通过可解释接口，一个在路口的自动驾驶车辆直行，能够给出是因为交通灯的状态是绿灯，刹车是因为前方检测到行人、障碍物或者检测到限速标志等原因。可解释性不仅对于驾驶者而言能够增强对自动驾驶系统的信任度，而且能够帮助自动驾驶系统的开发者对自动驾驶系统进行有效地设计、调试和诊断。可解释性还使得自动驾驶行为是可预测、可验证和可审计的。对于政府和政策制定者而言，可解释性还可以促使自动驾驶系统的合规性。

1.2　可解释交互的三个阶段

可解释性作为人与模型之间一种接口，是人与机器决策模型之间交流与协作的重要内容。解释的三个阶段见图 2，其大致可分为解释的生成、解释的传达和解释的接受[8]。

1）解释的生成。这一阶段主要是分析和提取无人车的感知和决策等模型的内在工作机理与决策逻辑等可解释的内容，从而帮助人们理解模型从数据中学到了什么，是如何决策的，决策行为的原因以及决策是否合理和可靠等。

图 1　自动驾驶的可解释性系统
Fig.1　Explainable autonomous driving system

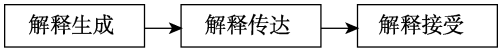

图 2 解释的三个阶段
Fig.2 Three phases of explanation

2）解释的传达。这一阶段是将解释的内容进行表达和呈现，主要涉及到内容呈现的粒度、时机、视角、通道形式以及具体的交互设计内容。

3）解释的接受。这一阶段主要研究用户在多大程度上能够理解和接受所呈现出来的解释内容，需要对解释进行多维度的、主观和客观的评价与用户测试。

1.3 可解释性的方法

目前的自动驾驶系统中感知和决策主要依赖于机器学习和深度学习算法，因而需要发现机器学习和深度学习模型背后的工作机理和决策逻辑，从中提取与生成可解释的内容。

1.3.1 具有自解释性的机器学习模型

传统机器学习模型大多本身就具有可解释性，例如最近邻（K-nearest Neighbors，KNN）、决策树（Decision Tree）、线性模型（Linear Model）、广义加性模型（Generalized Additive Model, GAM）和稀疏表示（Sparse Representation）模型等，其解释性主要体现在能够给出特征对决策的重要性度量。传统方法的解释性需要输入特征，其本身就具有一定的物理或者语义含义，由于模型准确度不够高，可解释性和模型性能往往存在一定的矛盾。对于深度神经网络，尽管其性能大幅超越了传统的方法，但是其内在工作机理并不透明，在应用中主要依靠工程经验和调参技巧。神经网络本身易受对抗样本的攻击，鲁棒性不足，存在较大的安全风险。

1.3.2 可视化分析方法

对于像深度神经网络这样高度非线性的机器学习模型，人们很难理解深度网络中隐藏层从数据中学到了什么，其背后的决策逻辑是什么。通过可视化的手段将任意隐含神经元计算内容进行可视化，以此捕捉神经网络中内部神经元计算内容的特定含义。例如激活最大化方法（Activation Maximization，AM）是一类典型方法，即寻找最大化激活给定的隐藏单元或者重构满足一定条件的输入模式[9]。AM 方法是一种全局解释方法。对于单个样本的局部解释方法，主要是针对每一个特定输入样本，通过分析和提取输入样本的每一维特征对模型最终决策的贡献程度，即提取特征的决策重要性，通过可视化手段进行呈现，使用户能从语义和视觉上直观理解模型对输入样本的决策逻辑和依据。例如基于梯度反向传播的系列方法，GuidedBP[10]、IntegratedGrad[11]和 SmoothGrad[12]等，利用神经网络的反向传播机制将对决策的重要性信号从模型的输出层逐层传播到模型的输入层，以推导输入样本的特征重要性，生成与之对应的决策显著性热力图（Heatmap），在输入图像中对决策的重要部分进行标注和显示。Zhou 等人提出了类激活映射（Class Activation Mapping，CAM）方法，利用全局平均池化（Global Average Pooling，GAP）层来替代传统 CNN 模型中除 Softmax 层以外的所有全连接层，并通过将输出层的权重投影到卷积特征图来定位图像中的重要区域[12]。Selvaraju 则将基于梯度的方法与 CAM 方法结合，提出梯度加权类激活映射方法（Grad-CAM）[13]。

1.3.3 基于代理模型的方法

基于代理模型的方法主要是用具有自解释的模型，例如规则列表、决策树和线性模型等，来近似原始黑盒模型的决策模型。例如，BozO 等人利用可解释的规则集合从神经网络模型中提取决策规则[13]，Chen 等人提出利用梯度提升树进行知识蒸馏的方式来学习可解释模型[14]。针对基于卷积神经网络 CNN（Convolutional Neural Network，CNN）的图像分类任务，Zhang 等人提出了基于 And-Or 图模型来解释 CNN 卷积层特征内在的图像知识层次，进而提取决策树规则来揭示卷积层中哪些滤波器会参与预测以及这些滤波器对预测结果的贡献程度[15]。Ribeiro 等人提出的 LIME（Local Interpretable Model-Agnostic Explanations）模型，其在一个样本邻域内，用一个线性模型来近似原非线性神经网络模型，线性模型的权重可作为输入特征局部重要性的指示器[16]，并进一步提出了 ALIME（Anchors-free LIME）方法，用清晰的“如果—那么”（if-then）的逻辑规则来解释模型的决策行为[17]。基于代理模型的方法是一类与模型无关（Model-agnostic）的方法。

1.3.4 基于原型的解释方法

人们在做复杂决策的时候，有时候并不是通过仔细分析和计算，而是基于相似的经验进行类比得出结论。基于原型（Prototype）的解释是通过选择有代表性的或者关键样本，来解释模型的决策行为。比较典型的方法是 Percy Liang 等人提出的基于影响力函数（Influence function）的方法来选择对一个分类器决策起到重要作用的样本，并以此来评估决策的合理性。基于影响力函数的方法还可以用来构建对抗样本，评估训练集与测试集分布的一致性以及发现训练集中的标记错误样本等[18]。Kim 提出基于 MMD（Maximum Mean Discrepancy）的方法来同时选择数据集的原型样本和所谓的 Critic 样本，进一步提升解释性[19]。

2 人-无人车交互中的可解释交互框架

无人车作为一个典型的智能体或者机器人，在与人的交互过程中，其决策和行为应当对人是透明和可解释的。Lyons 建立了 HRI（Human-Robot Interaction）中的透明度模型，一方面是智能系统应当传达给人的

信息，另一方面是涉及智能系统应当向人传达其在多大程度上意识到与理解了人的意图，并分别在这两个层面通过多个具体的模型来表明可解释的内容[20]。Pokam 等人则基于 Lyons 的 HRI 透明度模型，进一步提出了与无人交互相关的原则[21]。Hois 等人指出了人与智能系统的交互中可解释性的三个关键内容：（1）智能系统的行为及其意图的可解释性；（2）智能系统决策机制的可解释性；（3）智能系统的潜在限制以及给定情况下失误概率的可解释性。通过这三个方面可解释内容的提取、表达和呈现，增强系统的透明度和用户的信任度[22]。基于 HRI 中的可解释性与透明度的研究，本文提出了人-无人车交互中的可解释交互模型框架，见图 3。该模型框架中包含了系统功能模型、任务/环境/用户状态模型、决策机制模型以及行为/协同分工模型四类。

2.1　系统功能模型

与其他智能系统一样，无人车智能体的自主程度和功能范围等都各有不同。在高等级的自动驾驶汽车中，设计运行域（Operational Design Domain，ODD）的概念被用来划定其特定功能的特定运行条件，超出设计运行条件时将寻求人类驾驶员对车辆进行控制[23]。对于特定的无人车智能系统来说，其 ODD 一般是处于稳定的状态，并且应在交互开始前向用户进行明确传达。然而在实际交互的过程中，由于情境和条件的动态变化，用户难以准确把握无人车智能系统的能力边界、适用条件以及发生失误的可能性等。系统功能模型即要求在人与无人车交互中，准确地呈现其各项功能的最大自主程度与功能激活需要满足的内外部条件等。对系统功能范围及其鲁棒性的明确阐释也是人与无人车可解释性交互的基础。

2.2　任务/环境/用户状态模型

无人车内外布置的多种传感器赋予其强大的感知能力，在此基础上无人车智能体可以对其感知到的信息进行理解，包括任务信息、环境信息以及用户状态信息。任务模型、环境模型和用户状态模型是无人车智能体对其“在多大程度上感知到并理解了当前情境”进行解释的三个子模型。例如，用户主动发出了一个任务指令，无人车智能体需要给出有效的反馈，解释其对任务目标的理解并指示任务内容。无人车对动态环境实时的感知与理解，也需要在特定的任务情境下以合适的方式向用户进行传达。由于人也属于情境的重要因素，现在很多无人车中也融入了对于人的

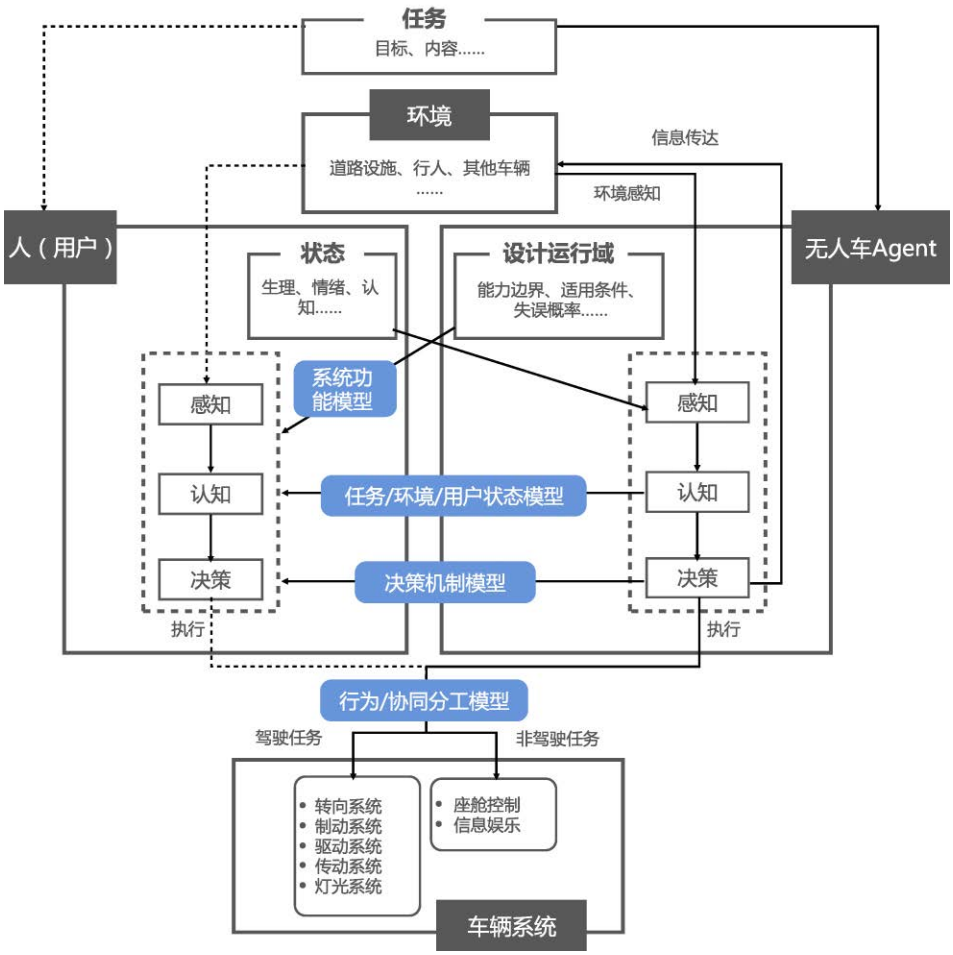

图 3　人-无人车交互中的可解释性交互框架

Fig.3　Framework of explainable human-machine interaction

生理、情绪和认知等要素水平的监测，这些因素会成为无人车智能体进一步决策的依据，而且需要将这种理解和决策过程以合适的方式向用户传达。

2.3 决策机制模型

基于任务目标以及对于情境的感知与认知，无人车智能体将在其能力范围内做出针对特定任务的决策。从上文中可以看到，支持智能系统进行决策的情境变量输入是十分复杂的，而 AI 算法（特别是深度机器学习算法）又具有“黑盒”特性，其内部机制是不透明的。因此，对无人车智能体如何处理输入其中的海量数据以形成决策的机制进行解释，是人与无人车交互中具有挑战性的问题。尤其是在具有较大不确定性的情境下，无人车智能体需要运用目标用户可理解的方式，对在当前条件下做出特定决策的机制或原因做出解释，以保持用户对无人车的恰当的信任关系，或支持用户介入无人车的决策闭环中。

2.4 行为/协同分工模型

在具体的任务执行层面上，存在无人车智能体自主执行和人与无人车协同执行的情况。当无人车智能体自主执行任务时，无人车需要对其当前的行为意图、执行状态和潜在风险等给出明确的解释，以使用户对于当前的任务执行保持良好的态势认知水平，对无人车接下来的行为进行预测。在人与无人车协同执行任务的情况下，无人车需要对给定任务中的人和无人车的功能分配进行清晰合理的解释，并在具体的执行状态方面与用户保持良好的交流，以免造成任务执行中的权责冲突。

3 人-无人车交互中可解释性设计要素

在人-无人车交互中涉及到多个层次框架，每个层次需要不同的解释内容。在具体进行解释表示和呈现中，将从解释的对象、方式和评价等方面，来探讨人与无人车交互中的可解释性设计要素，人-无人车交互中的可解释性设计框架见图 4。

3.1 解释对象

可解释的内容及其呈现与人的角色、驾驶经验、认识水平和知识背景等因素密切相关。对人-无人车交互进行可解释性设计，首先要确定解释所要面向的目标用户。在人与无人车交互的语境下，涉及的用户类型主要包括处于驾驶或监控位置的驾驶（安全）员、处于其他位置的乘客和车外的交通参与者等。不同角色的用户参与交互的目的和程度不同，个体之间的驾驶经验和对无人车系统的熟悉程度存在差异，对解释呈现的诉求也不同。驾驶（安全）员通常需要承担较多的驾驶任务，需解释的内容将覆盖从环境感知、规划决策到具体行为的多个层次。乘客通常作为服务的接受者，其与无人车的交互中非驾驶任务占比较大，需要解释的内容集中于策略性内容。面向车外交通参与者的解释更是强调无人车的短时行为及意图层面。

可解释还与具体驾驶场景有关。可解释性设计需要置入到不同的驾驶场景中。场景的核心是其中的事件，人与无人车构成的协同系统如何应对不同类型的事件，将涉及感知、决策与执行整个过程，最终产生不同的行为。因此，明确不同场景中的事件类型、性质和紧急程度，以及人与无人车的协同行为，能够为解释呈现方式的设计提供支持。

3.2 解释方式

在明确解释所面向的目标用户和场景的基础上，可以从解释的信息内容、呈现状态以及交互模态等方面进行设计考虑。更为详细的解释能够承载更多的信息，同时也将带来较大的认知负荷。因此，解释信息的详略程度应根据目标用户的认知特性与目标场景的关键程度等进行调整。另外，用户和无人车系统的相互了解程度也将在交互中随时间推移而加深，解释内容将逐渐从一般化趋于个性化。

在交互过程中，需呈现的解释信息丰富多样，各种解释信息的呈现时机与状态也十分重要。呈现模式包括以常驻显示的形式持续提供解释、以主动交互的形式由智能体择机提供解释和以被动响应的形式在用户寻求解释时做出反馈等。

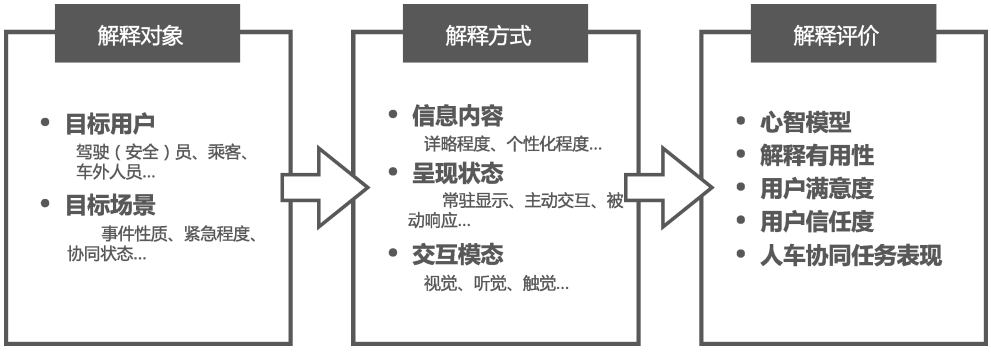

图 4 人-无人车交互中的可解释性设计框架

Fig.4 Design framework of explainable human-autonomous vehicle interaction

交互模态的选择往往取决于需要呈现的解释信息类型与内容。视觉是最重要的信息通道，既可以提供常态化的解释信息，也可以提供紧急场景下的高优先级解释信息，其具体呈现形式包括文字、图形、图像、动效与光效等。听觉通道通常承载优先级较高的解释信息，其具体呈现形式包括语音及声效等。触觉通道由于适用场景和可承载信息量有限，通常用于辅助解释信息的呈现。

3.3　解释评价

对于人-无人车交互中的可解释性设计评价，可以从用户的心智模型、解释的有用性、用户满意度、用户信任度以及人车协同的任务表现等方面来评价。心智模型是认知心理学领域的概念，表征用户如何理解一个系统。在人与无人车交互的语境下，解释将有助于用户建立对于无人车智能体如何运作的心智模型，因此研究用户的心智模型将有助于验证解释的有效性[24]。

解释的有用性、用户满意度以及信任度对于人-无人车交互中的用户体验十分重要。通过访谈与自我报告等定性方法以及问卷量表等定量方法，评测用户对于解释信息的投入度、认知负荷以及对智能系统的依赖程度等。人-无人车交互中解释的一个关键目标即是提升人车协同的任务绩效，往往通过设置特定的任务，测量任务完成的时间与准确率等指标来体现解释在多大程度上促进了人车协同的任务绩效。

3.4　典型案例分析

本文以蔚来的 NIO Pilot 自动辅助驾驶系统中的人机交互设计为例，分析其中的可解释交互[25]。NIO Pilot 开启下的解释信息界面见图 5，在可解释性交互框架的视角下，所呈现的主要信息可归到以下四个模型中：（1）系统功能模型，包括系统的模式与状态显示，指示系统具有的功能以及适合运行的条件，仅在车周围的状态显示环为蓝色时系统才可正常开启；（2）任务/环境/用户状态模型，以可视化形式呈现车辆感知到环境中的车道线、前方车辆与跟车时距等，以及监测到驾驶员未握住方向盘的状态；（3）决策机制模型，基于车辆感知信息所作出的正常跟车自适应巡航决策，以系统状态及跟车时距显示的部件色彩及动态等指示车辆行为；（4）行为及协同分工模型，呈现当前车辆的运行情况，如实时车速及设定巡航车速，以及需要驾驶员保持握住方向盘的任务请求。

从以上分析可以看到，解释信息的最终呈现与解释模型并非一一对应，一个解释元素往往是多个模型综合作用的结果。因此在实际的可解释性交互设计中，应当根据目标场景综合规划解释的内容与方式。NIO Pilot 变道响应场景下的可解释交互见图 6，在 NIO Pilot 系统对驾驶员主动发起的变道请求进行响应的场景下，首先对用户输入的任务进行反馈，以白

图 5　NIO Pilot 开启下的解释信息界面

Fig.5　Explainable information interface of NIO Pilot

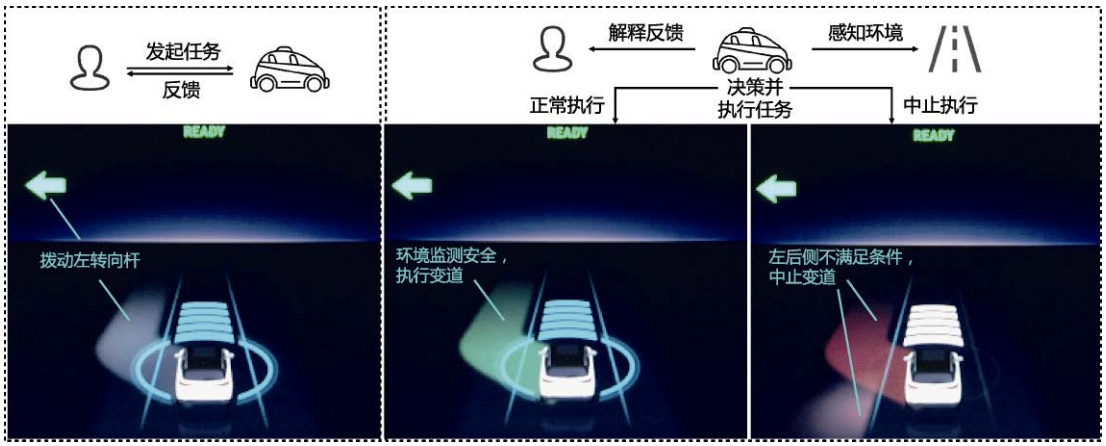

图 6　NIO Pilot 变道响应场景下的可解释交互

Fig.6　Explainable interaction in response to lane change request of NIO Pilot

色的变道指示图形表明将要执行向左变道的任务。车辆随即执行环境监测以判断是否满足变道条件，其决策以变道指示图形的色彩体现，绿色表示条件满足并正常执行变道任务，红色表示条件不满足并中止变道任务，并且进一步以动态图形解释其中止原因为左后方存在障碍物。该场景下的解释信息是随任务进展动态变化的，并且在出现任务失败时向用户进行主动的解释呈现。

4 结语

深度学习等现代人工智能技术已经广泛应用于自动驾驶系统中，其透明性和可解释性问题日益引起广泛关注。系统不仅决策如何行动，而且还需要提供如此决策的原因，并通过让人理解的方式进行表达和呈现。本研究旨在探讨人与无人车交互过程中为什么需要可解释性，需要什么样的可解释性内容，如何提取、呈现和评价这些可解释内容，以提高自动驾驶系统决策的透明度、安全性和用户信任度，建立无人车交互中的可解释交互框架，讨论可解释性交互设计要素。未来将结合具体的无人车驾驶场景，展开具体的可解释性内容研究及其交互设计实践，并基于自动驾驶仿真器和 VR/AR/MR 等测试平台，进行人-无人车交互中可解释性交互的测试。

参考文献：

[1] BROGGI A. Extensive Tests of Autonomous Driving Technologies[J]. IEEE Trans. Intell. Transp. Syst, 2013, 14(3): 1403-1415.

[2] ADADI A, BERRADA M. Peeking Inside the Black-Box: a Survey on Explainable Artificial Intelligence (XAI)[J]. IEEE Access, 2018, (6): 52138-52160.

[3] CYSNEIROS L M, RAFFI M. Sampaio Do Prado Leite. Software Transparency as a Key Requirement for Self-Driving Cars[C]. 2018 IEEE 26th International Requirements Engineering Conference (RE), 2018.

[4] WANG D, YANG Q, ABDUL A, et al. Designing Theory-Driven User-Centric Explainable AI[C]. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019.

[5] G D. Explainable Artificial Intelligence (XAI)[C]. DARPA/I2O, 2017.

[6] 纪守领, 李进锋, 杜天宇. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071-2096.
JI Shou-ling, LI Jin-feng, DU Tian-yu. Survey on Techniques, Applications and Security of Machine Learning Interpretability[J]. Journal of Computer Research and Development, 2019, 56(10): 2071-2096.

[7] KOO J, KWAC J, JU W, et al. Why Did My Car Just Do That? Explaining Semi-Autonomous Driving Actions to Improve Driver Understanding, Trust and Performanc[C]. 2015.

[8] ANJOMSHOAE S, NAJJAR A, CALVARESI D, et al. Explainable Agents and Robots: Results from a Systematic Literature Review[C]. 18th International Conference on Autonomous Agents and Multiagent Systems, 2019.

[9] NGUYEN A, CLUNE J, BENGIO Y, et al. Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space[J]. ArXiV, 2016, (11): 112.

[10] SPRINGENBERG J T, DOSOVITSKIY A, BROX T, et al. Striving for Simplicity: the All Convolutional Net[J]. ArXiV, 2014, (12): 98.

[11] SUNDARARAJAN M, TALY A, YAN Q. Gradients of Counterfactuals[J]. ArXiV, 2016.

[12] SMILKOV D, THORAT N, KIM B, et al. SmoothGrad: Removing Noise by Adding Noise[J]. arXiv, 2017, (11): 324.

[13] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual Explanations from Deep Networks Via Gradient-Based Localization[J]. Int. J. Comput. Vis., 2020, 128(2): 336-359.

[14] C Z, P S, K R. Interpretable Deep Models for ICU Outcome Prediction[C]. AMIA Annu Symp Proc, 2017.

[15] ZHANG Q, YANG Y, MA H, et al. Interpreting CNNs Via Decision Trees[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[16] RIBEIRO M, SINGH S, GUESTRIN C. Why Should I Trust You? Explaining the Predictions of Any Classifier[C]. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics, 2016.

[17] RIBEIRO M T, SINGH S, GUESTRIN C. Nothing Else Matters: Model-Agnostic Explanations by Identifying Prediction Invariance[J]. ArXiV, 2016, (11): 118.

[18] K P W, L P. Understanding Black-Box Predictions Via Influence Functions[C]. International Conference on Machine Learning, 2017.

[19] KIM B, KHANNA R, KOYEJO O O. Examples are Not Enough, Learn to Criticize! Criticism for Interpretability[C]. Advances in Neural Information Processing Systems 29, 2016.

[20] LYONS J B. Being Transparent about Transparency: a Model for Human-Robot Interaction[C]. AAAI Spring Symposium Series, 2013.

[21] POKAM R, DEBERNARD S, CHAUVIN C, et al. Principles of Transparency for Autonomous Vehicles: First Results of an Experiment Ith an Augmented Reality Human-Machine Interface[J]. Cogn. Technol. Work, 2019, 21(4): 643-656.

[22] HOIS J D, THEOFANOU F, JUNK A J. How to Achieve Explainability and Transparency in Human AI Interaction[C]. HCII 2019 Communications in Computer and Information Science, 2019.

[23] FARAH H. An Empirical Analysis to Assess the Operational Design Domain of Lane Keeping System Equipped Vehicles Combining Objective and Subjective Risk Measures[J]. IEEE Trans, 2020, (4): 1-10.

[24] MOHSENI S, ZAREI N, RAGAN E D. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems[J]. ArXiV, 2018, (11): 24.

[25] 章建勇. NIO Pilot 主要功能干货指南(上篇)[EB/OL]. (2019-06-10)[2020-06-28]. https://www.nio.cn/app-article-503822.
ZHANG Jian-yong. Guide of Main Functions of NIO Pilot (I) [EB/OL]. (2019-06-10)[2020-06-28]. https://www.nio.cn/app-article-503822.