精确性越高，效果反而越差：人们能否识别基于概念的可解释人工智能中的不足场景表征?

Romy Müller & Wiebke Klausing

工程心理学与应用认知研究系，德累斯顿工业大学

摘要

可解释人工智能（可解释人工智能）旨在帮助揭示人工智能模型内部表征中的缺陷。但人们能否从其解释中得出正确结论？具体而言，他们能否识别人工智能无法区分相关特征与无关特征？本研究中，模拟人工智能将铁路擅入者的图像分类为危险或不危险。为解释人工智能使用了哪些特征，研究向参与者展示了数据集中其他能够以相似方式激活人工智能的图像。这些概念图像在三个相关特征（即人物与轨道的距离、方向和动作）以及一个无关特征（即场景背景）上有所不同。当人工智能在决策中使用某项特征时，该特征会保留在概念图像中；否则，图像会在该特征上随机变化（例如，距离相同、背景不同）。当概念图像保留了相关特征时，参与者对人工智能的评价更高。对于无关特征，他们通常并不介意，有时甚至更希望该特征得到保留。这表明，人们可能无法识别人工智能模型在决策时依赖无关特征的情况。

关键词：可解释人工智能；基于概念的可解释人工智能；类别表征；泛化；人工智能偏差

引言

可解释人工智能（可解释人工智能）旨在帮助人们理解人工智能模型如何做出决策。例如，为了解释图像中描绘的情境为何被归类为危险，可以展示人工智能使用了图像中的哪些区域，或者它在那里看到了什么。此类可解释人工智能不仅在实践层面前景广阔，也引出了关于人类认知的有趣问题。一个值得探讨的问题是，人们如何利用不一致之处来推断其他主体（人类或人工智能）是如何表征类别的。例如，当他人始终如一或前后不一地使用某些特征时，这说明他们如何理解某个类别?

设想一个人工智能图像分类器判定某种情境是危险的。为了解释原因，系统会展示若干其他图像，这些图像以相同方式激活其神经元。你大概会希望所有这些图像都描绘相似的危险来源（例如，有人靠近铁路轨道，同时列车正在驶近）。现在设想所有这些图像还显示了相同的场景背景。你会对此作出积极评价，认为这表明人工智能已形成对场景的精确表征吗？还是会认为它缺乏充分的场景理解，因为它使用无关特征来判定危险？这可能会让你担心人工智能是否已形成足够鲁棒的危险表征，从而无法在其他情境中同样识别出危险。

先前的一项研究表明，人们希望人工智能模型尽可能精确地表征场景，却未意识到这可能给人工智能的鲁棒性带来潜在代价（Müller, 2025b）。具体而言，当用于解释的概念图像紧密复现被分类场景时，参与者对人工智能性能的评分高于这些图像仅对一个较不相关的特征进行泛化时。本研究更系统地考察了这种对精确性的偏好。也就是说，我们探究了人们希望或不希望在基于概念的可解释人工智能中保留哪些图像特征。在描述本研究之前，我们将讨论人工智能依赖无关特征的问题、可解释人工智能揭示此类偏差的潜力，以及人类可能如何回应这些解释。

精确性与鲁棒性之间的权衡

显然，人工智能模型，尤其是图像分类器，应当准确表征其接收的输入。同时，它们的表征不应过度精确：其中只应包含真正相关的特征。然而，表征无关特征是人工智能中的常见问题。首先，除相关特征外，表征中还可能包含无关特征。这正是过拟合的情形（Ying, 2019）。当人工智能模型经过过度训练，因而形成过度精确的表征时，就可能出现这种情况。由此产生的问题是，情境发生变化时，人工智能将无法再做出正确决策。图像分类器因缺乏泛化能力而鲁棒性低下，是一个严重的问题（Muttenthaler等人，2024）。

第二个问题出现在人工智能使用无关特征而非相关特征时，这通常源于数据集偏差（Torralba和Efros，2011年）。当相关特征和无关特征在数据集中具有相关性时（例如，所有危险图像都是在同一地点拍摄的），人工智能可能并未真正学习相关特征（例如，靠近轨道的人），而是将具有相关性的特征作为捷径。此类捷径学习是图像分类中的常见现象（Geirhos等人，2020年）。一种普遍存在的偏差是，人工智能模型依赖场景背景，因此当背景发生变化时便无法再对图像进行分类（Beery等人，2018年）。例如，它们无法识别海滩上的牛；或者当画面中实际上只有绿色草地时，却错误地识别出羊。

虽然过拟合与捷径学习的成因不同（例如，训练过多与数据集偏差），但一个共同主题是：人工智能依赖于无关特征。它要么在相关特征之外还使用这些特征（过拟合），要么用它们取代相关特征（捷径学习）。两者都可能导致缺乏

鲁棒性：人工智能在不共享相关特征与无关特征之间相同关联的新情境中将无法正常工作。因此，这些问题应当被检测出来，而可解释人工智能旨在促进这种检测。

基于概念的可解释人工智能能否揭示过度精确的人工智能表征？

人工智能在图像中看到了什么？基于概念的可解释人工智能可以展示人工智能用于做出决策的概念，从而让这一过程变得透明（综述参见 Poeta等人，2023）。一种方法是展示数据集中其他能够以类似于分类图像的方式激活人工智能的图像或图像块（Achtibat等人，2023；Fel等人，2023）。实证研究表明，这种基于概念的可解释人工智能确实有助于人们发现人工智能偏差（Achtibat等人，2023；Adebayo等人，2022；Colin等人，2022；Qi等人，2021）。因此，人们更能识别出人工智能模型何时依赖无关特征，例如场景背景或其他数据伪影。

然而，在这些研究中，可解释人工智能通常呈现的是单一特征或个体特征（例如，放射影像上的水印）。这些特征很容易在概念图像中被单独呈现，使人类评估者将注意力集中在一个包含所有必要信息的小图像块上。这让人们相当容易识别人工智能使用了哪些特征，也容易得出人工智能不应使用这些特征的结论。相反，对危险这类复杂类别的分类往往依赖非单一特征或关系特征（例如，某人与铁路轨道之间的距离），也依赖多特征组合（例如，距离和动作）。在这种情况下，概念图像必然会同时包含多个特征，人类评估者需要并行处理这些信息。例如，这类概念图像可能展示一个身处场景中的人，而非单独呈现一块背景。如果背景对人工智能决策有所贡献，这一点就会体现在它的重复出现上：同一背景会出现在分类图像中，也会出现在所有概念图像中。想必这会让人工智能偏差更难被发现，因为人们需要从多张图像中无关特征的重复出现推断出偏差。此外，人们还必须得出结论，认为这种重复是不妥的。换句话说，他们必须因为人工智能过度精确而惩罚它：人工智能纳入了无关特征，而没有对该特征进行随机化处理。下一节将讨论我们是否可以预期人们会这样做。

人们会欣赏较低的精确性吗？人们是否会欣赏较低的精确性

人们对可解释人工智能概念的判断，很可能取决于他们如何解读这种情况：是将其视为人工智能的缺陷，还是将其视为人工智能具备泛化能力的迹象。这并非无关紧要，因为人类和人工智能的泛化方式并不相同（Ilievski等人，2025；Thalmann与Schulz，2025）。因此，当人工智能无法形成足够一般的表征时，人们甚至可能意识不到这一点。事实上，一项先前的研究表明，人们偏好精确的表征，而非概括，并且无法将后者与系统性错误区分开来（Müller，2025b）。例如，当分类图像中显示一个人在铁路轨道上行走时，可解释人工智能通过展示人们在铁路轨道上从事各种活动的图像，给出了一个更一般的概念，例如“在轨道上”。与保留分类图像中动作的精确信息相比，这使参与者对人工智能表现的评价更低；而这一评价与系统性错误的情况同样低，后者的所有图像都显示了错误的动作（例如，站立而非行走）。不过，有人可能会认为，在评估危险时，对动作进行概括可能会带来问题，因此动作并非真正的无关特征。这就引出了一个问题：特征的相关性如何影响人们对可解释人工智能中不精确性的感知？如果人工智能的不精确表征是对完全无关的特征进行概括，人们会不会对这些表征更为认可？

本研究

图像分类器的概念表征不应包含无关特征。这也意味着，基于概念的可解释人工智能只会保留相关特征，并对无关特征进行随机化。人们会欣赏由此产生的不精确性吗？他们能否认识到，一定程度的变化有利于提升鲁棒性？本研究在铁路安全场景中探讨了这些问题。参与者观看了铁路轨道附近有擅闯者的静态图像，并评估了一个模拟人工智能，该系统已判定每种情形是否危险。系统通过展示数据集中其他图像来解释这些决策；根据人工智能的判断，这些图像在影响其决策的特征方面与当前图像相似。

从人的角度来看，其中三个特征是相关的：人物与轨道的距离、人物相对于轨道的朝向，以及人物的动作。不过，这些特征的相关性可能有所不同。首先，研究一再发现，距离对人们评估人工智能性能的影响最大（Müller，2025a, 2025b）。其次，目前尚无关于方向的实证数据，但方向可以预测是否可能发生碰撞，因此其相关性或许与距离大致相当。第三，动作被视为较不重要的特征，但参与者仍然更喜欢人工智能精确地表征动作（Müller，2025b）。最后，还有一个明显无关的第四项特征：场景背景。在实际铁路交通中，环境是预测危险的重要因素，列车司机可以据此作出各种推断（Müller & Schmidt，2025）。然而，非专业参与者大概无法做到这一点。更重要的是，此处使用的图像均拍摄于铁路测试设施，因此其背景仅在表面特征上有所不同。

关键的实验操控是：概念图像中的哪些特征保持相同，哪些特征有所变化。我们预期总体上会存在相同性偏好，但假设这种偏好会受到特征相关性的调节。也就是说，参与者对相关性较高的特征应表现出更强的相同性偏好（即距离、方向、动作，依次递减）。然而，最有趣的结果涉及无关特征（即背景）。基本上，可能出现三种结果。

首先，如果参与者遵循规范行事，那么他们的偏好应该会反转：背景有所变化时的评分应高于背景保持相同时的评分。鉴于人们似乎偏好精确性（Müller, 2025b），我们认为这种结果不太可能出现。其次，如果参与者并不在意是否保留无关特征，那么背景相同与不同情况下的评分应无差异。第三，如果他们希望人工智能尽可能精确地表征场景，那么他们甚至可能对背景表现出相同性偏好。

方法

参与者

我们从该大学的参与者招募池中招募了59名参与者，他们获得了部分课程学分或每小时10欧元的报酬。参与者年龄介于19至61岁之间（M = 28.1，SD = 9.0），其中41人为女性，18人为男性。

实验设备与刺激材料

实验室设置 最多四名参与者在安静的实验室房间内同时进行实验。指导说明和实验程序在四台台式电脑（24英寸）中的一台上运行，并使用标准电脑鼠标作为输入设备。

指导说明 指导视频（5分30秒）以配有音频旁白的 Microsoft PowerPoint 演示文稿为基础。视频介绍了铁路场景，并解释了基于概念的可解释人工智能的原理（即呈现数据集中能以与分类图像相似的方式激活人工智能的图像）。视频以识别狗这一玩具示例为例，告知参与者，概念图像可能与分类图像恰当匹配、差异过大、过于相似，或不适合用于决策。最重要的是，视频明确告知参与者，过高的相似性会引出一个问题：人工智能是否足够鲁棒，能够识别其他图像。此外，视频还告知参与者，对于危险这类复杂概念，重要的不仅是图像中有哪些物体，还有场景中正在发生什么。最后，视频通过实验的屏幕截图说明了实验流程。

刺激材料 所有刺激材料的分辨率均为 1920 × 1080 像素，包含图像、文本（德语）和交互元素。参与者会看到两个屏幕。决策屏幕显示分类图像（650 × 225 像素）、人工智能决策（即以红色字体显示“危险”，或以绿色字体显示“不危险”），以及两个按钮，分别用于指示该决策是否正确。人工智能评分屏幕（见图1）再次显示分类图像和人工智能决策，但还增加了五张纵向排列的概念图像（每张 433 × 150 像素）以及一个滑块，用于指示人工智能的表现如何（采用 100 分制，范围从“非常差”到“非常好”）。



图1：展示人工智能评分屏幕的刺激示例。

共有 128 种不同的刺激材料，每种都由一张分类图像、一项人工智能决策和五张概念图像组成。在这些刺激材料中，分类图像与概念图像之间的匹配关系在四个特征上进行了因子操纵：人物与轨道的距离、人物转身的方向、人物执行的动作以及场景背景。在每个特征上，概念图像可能与分类图像相同，也可能有所不同。某个特征相同，意味着五张概念图像中的该特征都与分类图像相对应（例如，所有图像都显示远离轨道的人）。相反，某个特征有所变化，意味着五张概念图像中只有 2-3 张在该特征上与分类图像相对应，其余 2-3 张则有所不同。

距离这一特征进一步细分为两类刺激材料：人物靠近轨道，或距离轨道较远（分别为小于或大于1米）。同样，方向被细分为人物朝向轨道或背向轨道的刺激材料，而动作则被细分为人物正在行走或执行其他动作（例如站立、坐着、跪着）的刺激材料。当动作与分类图像中的动作相同时，始终为行走（示例见图2）。

实验流程

实验开始前，参与者观看了指导视频。本实验采用 2（距离）× 2 （方向）× 2（动作）× 2 （背景）被试内设计。每个因素分别有“相同”和“不同”两个水平。实验共包含 128 个试次，由 64 个刺激材料分别与“危险”和“不危险”两种人工智能决策组合而成。每位参与者的试次顺序均经过随机化。在每个试次的第一阶段，参与者只会看到分类图像和人工智能决策。他们通过按下相应按钮，指出人工智能决策是正确还是错误。按下按钮后，试次进入第二阶段，按钮消失，取而代之的是概念图像和滑块。参与者可以点击滑块上 0 至 100 之间的任意位置，为人工智能的表现评分。整个实验耗时约 1 至 1.5 小时。

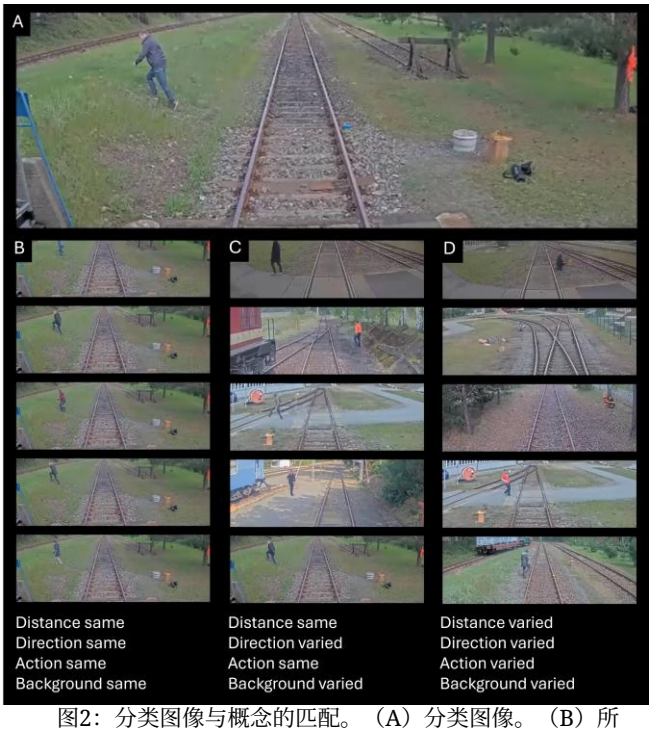


图2：分类图像与概念的匹配。（A）分类图像。（B）所有特征均相同，（C）一半特征相同。（D）所有特征均有变化。

结果

人工智能性能评分人们

’对人工智能的评分 ’的性能

表现很大程度上取决于他们是否认同其决策（Müller, 2025b）。因此，分析中将一致性作为一个因素。我们使用 2（一致性）× 2 （背景）× 2 （动作）× 2（方向）× 2 （距离）重复测量方差分析，对人工智能性能的平均评分进行分析。所有两两比较均采用邦弗伦尼校正。对于更高阶交互作用，我们有选择地从无关特征（即背景）的角度报告效应，从而检验参与者对相同背景与不同背景的偏好是否取决于其他三个特征。

对于三个相关特征，显著的主效应表明，参与者偏好相同图像而非变化图像（见图3）。距离方面（60.1 vs. 48.5），$F(1,58) = 50.436$，$p < .001$, $\eta_p2 = .465$；方向方面（58.4 vs. 50.2），$F(1,58) = 39.437$，$p < .001$, $\eta_p2 = .405$；动作方面（57.4 vs. 51.2），$F(1,58) = 50.519$，$p < .001$, $\eta_p2 = .466$。数值上，相同性偏好在距离方面最大，在方向方面居中，在动作方面最小。相反，背景没有主效应，$F(1,58) = 2.800$，$p = .100$, $\eta_p2 = .046$，表明参与者对相同与变化图像的评分没有差异（55.0和53.6）。

此外，对于这两个关系特征，当参与者与人工智能的判断一致时，相同性带来的益处更大；方向与一致性之间存在显著交互作用，$F(1,58) = 19.102$, $p < .001$, $\eta_p2 = .248$, 距离以及一致性，$F(1,58) = 4.460$，$p = .039$，$\eta_p2 = .071$；此外，方向、距离和一致性 $F(1,58) = 5.281$，$p = .025$，$\eta_p2 = .083$。

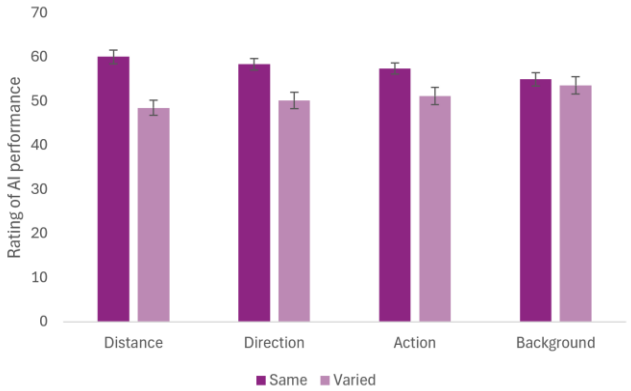


图3：各特征的人工智能性能评分。误差线表示均值标准误。

部分特征之间也存在交互作用。首先，距离与方向之间存在交互作用，$F(1,58) = 7.420$, $p = .009$, $\eta_p2 = .113$；方向与动作之间也存在交互作用，$F(1,58) = 12.692$, $p < .001$, $\eta_p2 = .180$。在这两种情况下，两两比较都呈现出一致的模式：尽管每项比较中的相同性偏好均达到显著水平，所有 p 均 < .001，但当另一项相应特征也相同时，某一特定特征的相同性偏好更强。此外，背景评分取决于距离，$F(1,58) = 6.353$, $p = .014$, $\eta_p2 = .099$。当距离相同时，相同背景与不同背景的评分相近，$p = .899$；而当距离不同时，相同背景的评分高于不同背景，$p = .010$。因此，当最相关特征不一致时，参与者偏好无关特征保持一致。

在更高阶的交互作用方面，背景、动作和距离之间存在三重交互作用，$F(1,58) = 4.141$，$p = .046$，$\eta_p2 = .067$。这表明，当动作相同但距离不同时，参与者偏好背景相同，$p < .001$；但在所有其他组合中，他们对背景没有偏好，所有 ps > .2。最后，背景、动作、方向和距离之间存在四重交互作用，$F(1,58) = 4.438$，$p = .039$，$\eta_p2 = .071$（见图4）。两两比较检验了在其他特征的每种组合下，参与者是否偏好相同背景。结果显示，当距离不同时，如果动作和方向都相同，参与者显著偏好背景相同，$p = .043$；或者，当只有动作相同、而距离和方向这两个关系特征不同时，参与者也显著偏好背景相同，$p < .001$。对于所有其他组合，背景相同或变化并无差别，所有 ps > .07。最值得注意的是，当所有相关特征都相同时，背景是否相同并不重要，$p = .696$。综上，这些更高阶的交互作用表明，参与者有时偏好相同背景，尤其是在最相关的特征不同时。

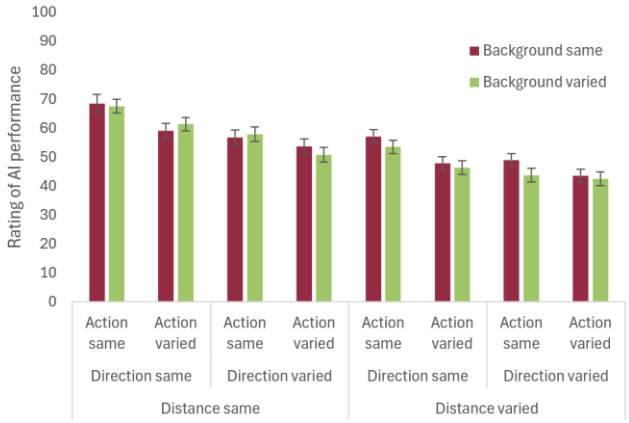


图4：其他三个特征的所有组合下的背景相同性效应。误差线表示均值标准误。

相关性与个体间差异

我们进行了两项探索性分析，以更好地了解参与者的评分策略。首先，我们询问不同特征的相同性偏好是否相互依赖。为检验这一点，我们计算了每种特征在相同与变化图像中的评分差值之间的相关性。方向和动作之间存在较高且显著的相关性，$r = .605$，$p < .001$。因此，喜欢展示相同方向的图像的参与者，也喜欢展示相同动作的图像。其他特征之间的相关性均不显著，所有 $rs < .3$，所有 $ps > .08$。

第二项探索性分析聚焦于个别参与者对背景的评分。对于这一无关特征，先前并未发现相同与变化的图像之间存在总体差异。这是因为参与者不在意背景，还是因为有些人偏好背景相同，而另一些人偏好背景变化，导致相反的偏好相互抵消？为检验这一点，我们统计了对相同与变化的图像的评分差异至少达到+/- 5个评分单位（满分为100分）的参与者人数。绝大多数参与者（40人，占67.8%）没有表现出这种差异；14人（占23.7%）偏好背景相同，只有5人（占8.5%）偏好背景变化。因此，大多数参与者似乎并不在意人工智能在其表征中是否纳入背景。

讨论

人们是否总是希望人工智能系统在其表征中尽可能精确，还是会为了更高的鲁棒性而欣赏一定的不精确性？若是后者，那么当基于概念的可解释人工智能解释在无关特征上发生变化时，他们应当更偏好这种解释。这种变化将表明人工智能有能力超越特定情境的细节进行泛化。为评估不精确性如何影响对人工智能的评价，参与者必须在铁路安全场景中对人工智能性能进行评分。用于解释人工智能决策的概念图像包含四个与危险目标概念相关性各不相同的特征。这些特征要么是与分类图像中的相同，要么在不同图像之间发生变化。从规范角度看，当所有概念图像都保留相关特征（即距离、方向、动作），但对无关特征（即背景）进行随机化时，参与者应当给人工智能更高的评分。然而，结果呈现出截然不同的图景。参与者总体上偏好相同性，而且这种偏好的程度会随着特征相关性的增加而增大。最重要的是，他们从未因无关特征的变化而给予奖励。这一结果令人担忧，将在以下各节中加以讨论。

无关特征的变化

总体而言，参与者对背景相同与变化的图像给出的评分并无差异。这是否意味着他们不在意人工智能是否利用背景来判断情境是否危险？未必。另一种可能是，参与者在评估人工智能性能时，回答的并非正确的问题。也就是说，他们或许没有评估人工智能的内部表征，而只是评估所呈现特征的相关性。如果某个特征与危险无关，他们就会忽略它。用一个较简单的问题替代复杂问题的倾向称为替代偏差，这种偏差在许多决策情境中都很常见（Kahneman 与 Frederick，2002）。在评估人员检测人工智能系统性能的情境中，也观察到了类似问题（Müller，2025a）。参与者被明确告知，该人工智能系统仅用于检测人员，无法评估情境是否危险。尽管如此，当系统漏检处于危险中的人员时，参与者仍会给它较低的性能评分。因此，人们评估人工智能时，可能不会关注人工智能表征了什么，以及它如何利用这些表征完成指定任务。相反，他们似乎依赖自己关于场景中哪些因素有助于做出良好决策的先验知识。显而易见，这种替代的风险在于，人们可能无法察觉人工智能缺陷，因而也不会采取措施加以纠正。

特征之间的交互作用

人们对场景背景的使用不足，是否取决于其他特征是否得到正确表征？首先，这可以表明他们是否对过拟合或捷径学习不敏感。此外，这或许有助于我们理解人们在评估人工智能时究竟在做什么。

关于第一个问题，参与者似乎对过拟合无动于衷，甚至鼓励捷径学习。如果人工智能出现过拟合，就会在使用相关特征的同时使用无关特征。两两比较结果表明，当所有相关特征都得到正确表征时，参与者对相同背景和不同背景的评分并无差异。因此，他们似乎既没有察觉过拟合，也没有因此给予奖励。相反，学会走捷径的人工智能会使用无关特征，而非相关特征。在这种情况下，参与者甚至会奖励人工智能不充分的表征。具体而言，当距离和方向发生变化时，参与者明显更偏好相同背景。

这或许能告诉我们，参与者的评估策略是什么？一种可能是，他们希望人工智能的表征至少有某些方面能够吻合。也就是说，当人工智能未能正确表征相关关系特征时，至少应该能够识别整体场景。另一种可能是，参与者将从人与人之间解释中获得的经验迁移了过来。人类与可解释人工智能在提供解释的方式上有所不同（Qi等人，2024）。例如，人们会有策略地在解释中着重说明某些方面，同时省略其他方面，从而让受众能够推断世界状态（Kirfel et al., 2022）。如果想突出相关差异，最好只改变关注特征，同时保持无关细节不变。事实上，在本研究中，当背景保持一致时，参与者的评分在相同方向和不同方向之间的差异更为明显。因此，无关特征的一致性或许可以视为一种解释性优点，因为它有助于人们聚焦于相关差异。不过，这只是高度推测性的说法；迄今为止，人们对某些特征保持一致的偏好如何以及为何取决于其他特征的状态，仍远未明了。

评估人工智能时的偏见真的能轻易消除吗？

有人可能会认为，本研究中观察到的参与者评估策略不足很容易纠正。毕竟，只要明确告诉他们，概念呈现相同背景是不好的就行了。然而，这样是否能彻底解决问题，似乎值得怀疑。首先，参与者实际上已经被指导说明，概念之间可能过于相似。这一点还通过一个例子加以说明，并明确指出了这对人工智能鲁棒性造成的风险。当然，我们总可以调整指导说明，确保参与者真正理解问题，并能将其运用到自身情境中。

更值得关注的是，在现实生活中，过度特定化可能很难被察觉。在本研究中，概念图像中特征的一致性被简化为一个二元变量（即相同／多样）。例如，当人工智能依赖背景时，所有图像中的背景都会完全重复，因此很容易察觉。而在更贴近现实的情境中，要意识到基于概念的可解释人工智能中的相似性过高，就困难得多。例如，这可能意味着可解释人工智能只展示来自乡村地区的概念图像，或只展示在晴朗天气下拍摄的图像。因此，具体场景从未重复出现，人们甚至很难注意到这些不理想的相似之处。从这个角度来看，参与者似乎连本研究中如此明显的过度特定化都不介意，这一点相当令人担忧。

局限与展望

在根据本研究得出结论之前，应考虑其若干局限性。从实践角度来看，存在几项对外部效度的威胁。我们只使用了静态图像而非动态场景、模拟的人工智能和可解释人工智能而非真实系统，并且使用的是新手参与者而非铁路安全专家或人工智能专家。所有这些因素都可能对结果产生影响。

此外，研究在内部效度方面也存在局限。其中之一是，我们没有将相同性推至极限。即使在保留了全部四项特征的试次中，仍有一些本文未考虑的其他特征存在差异，例如人物身份、服装或运动阶段。因此，我们实际上无法得出结论，认为参与者偏好人工智能的概念尽可能贴近分类图像。参与者的相同性偏好大概存在一定限度，甚至可能有临界点。未来研究应探讨：概念图像的相似性边界需要被突破到何种程度，参与者才会意识到变化是可取的。例如，所有图像中都可以出现同一个人，从而暗示人工智能依赖人物身份。若过度特定化进一步加剧，最终是否会使参与者的总体策略发生转变，让他们普遍更关注不受欢迎的重复现象，这一点也值得研究。

一个更具概念性的局限是，我们没有对比概括与系统性错误。变化的概念是指人工智能对某项特征进行随机化（例如，距离各不相同）所形成的概括。相反，研究中没有任何试次会在所有图像中都使用同一个错误特征（例如，所有人都远离轨道，而分类图像中的人则离轨道很近）。因此，我们无法判断不精确性的类型（即随机还是系统性的）会如何影响参与者的评分。先前一项研究发现，对于动作而言，两者并无差别（Müller, 2025b）。然而，当前研究结果表明，动作被视为相关性中等的特征。因此，未来研究应针对相关性高和低的特征，对比概括与系统性错误。

结论

人们在接收到基于概念的解释时，能否充分评估人工智能模型的内部表征？也就是说，他们是否能体会到有助于确保人工智能鲁棒性的有益不精确性？本研究表明，事实并非如此。参与者似乎并不在意基于概念的可解释人工智能是否暗示人工智能已将场景背景纳入其表征和决策之中。这或许表明，基于概念的解释无法让参与者对人工智能系统的内部表征得出有效结论。因此，未来研究一个令人振奋的方向是探索人类如何理解人工智能的概念表征。这最终或许能帮助我们更好地协调人类与人工智能代理的视角。

致谢

本研究得到联邦铁路局德国铁路交通研究中心（DZSF）“铁路安全评估的可解释人工智能（XRAISE）”项目的支持

参考文献

Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., & Lapuschkin, S.（2023）。通过概念相关性传播，从归因图到人类可理解的解释。《自然·机器智能》，5（9），1006-1019。https://doi.org/10.1038/s42256-023-00711-8

Adebayo, J., Muelly, M., Abelson, H., & Kim, B.（2022）。事后解释可能无法有效检测未知的虚假相关。在《国际学习表征会议》（第1-13页）。Beery, S., van Horn, G., & Perona, P.（2018）。未知领域中的识别。《欧洲计算机视觉会议论文集》，456-473。Colin, J., Fel, T., Cadène, R., & Serre, T.（2022）。我无法预测的，就无法理解：用于评估可解释性方法的以人为中心的框架。在《第36届神经信息处理系统大会》（第1-14页）。

Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., Colin, J.,...Serre, T.（2023）。CRAFT：用于可解释性的概念递归激活分解。在《IEEE/CVF计算机视觉与模式识别会议论文集》（第2711-2721页）。

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A.（2020）。深度神经网络中的捷径学习。《自然·机器智能》，2（11），665-673。https://doi.org/10.1038/s42256-020-00257-z Ilievski, F., Hammer, B., van Harmelen, F., Paassen, B., Saralajew, S., Schmid, U.,...Villmann, T.（2025）。协调人类与机器之间的泛化。《自然·机器智能》，7（9），1378-1389。https://doi.org/10.1038/s42256-025-01109-4

Kahneman, D., & Frederick, S.（2002）。重新审视代表性：直觉判断中的属性替代。在T. Gilovich、D. W. Griffin与D. Kahnemann（编）《启发式与偏差：直觉判断心理学》（第49-81页）。剑桥大学出版社。

Kirfel, L., Icard, T., & Gerstenberg, T.（2022）。基于解释的推断。《实验心理学杂志：通论》，151（7），1481-1501。https://doi.org/10.1037/xge0001151

Müller, R.（2025a）。人类如何评估用于列车自动运行的人员检测人工智能系统：并非所有漏检都相同。《未来交通》，5（3），78。https://doi.org/10.3390/futuretransp5030078

Müller, R.（2025b）。当基于概念的可解释人工智能不够精确时：人们能区分概括与失实表征吗？arXiv预印本，arXiv:2506.17936，1-23。https://doi.org/10.48550/arXiv.2506.17936

Müller, R.，& Schmidt, J.（2025）。超越物体识别：列车司机如何评估碰撞风险。《认知、技术与工作》，1-23。https://doi.org/10.1007/s10111-025-00832-7

Muttenthaler, L.，Greff, K.，Born, F.，Spitzer, B.，Kornblith, S.，Mozer, M. C.，……Lampinen, A. K.（2024）。跨抽象层级对齐机器与人类视觉表征。arXiv预印本，arXiv:2409.06509，1-54。https://doi.org/10.48550/arXiv.2409.06509 Poeta, E.，Ciravegna, G.，Pastor, E.，Cerquitelli, T.，& Baralis, E.（2023）。基于概念的可解释人工智能：综述。arXiv预印本，arXiv:2312.12936，1-46。https://doi.org/10.48550/arXiv.2312.12936 Qi, R.，Zheng, Y.，Yang, Y.，Cao, C. C.，& Hsiao, J. H.（2024）。人类与当前可解释人工智能中的解释策略：来自图像分类的洞见。《英国心理学杂志》，1-24。https://doi.org/10.1111/bjop.12714 Qi, Z.，Khorram, S.，& Fuxin, L.（2021）。将深度网络嵌入视觉解释。《人工智能》，292，103435。https://doi.org/10.1016/j.artint.2020.103435 Thalmann, M.，& Schulz, E.（2025）。我们如何刻画人类泛化，并将其与机器泛化区分开来？《心理科学最新进展》，34(5)，293-300。https://doi.org/10.1177/09637214251336212 Torralba, A.，& Efros, A. A.（2011年6月20日至25日）。对数据集偏差的无偏审视。载于《计算机视觉与模式识别会议》（第1521-1528页）。科罗拉多斯普林斯，科罗拉多州。电气电子工程师学会。Ying, X.（2019）。过拟合及其解决方案概述。《物理学杂志：会议系列》，1168，022022。https://doi.org/10.1088/1742- 6596/1168/2/022022