

教AI用更像人类的方式看世界

作者：Andrew Lampinen, Klaus Greff 2025年11月11日

【译文】 一项新研究表明，重组模型的视觉表征（visual representations）能使其（AI）更有用、更鲁棒、更可靠。

“视觉”人工智能（AI）无处不在。我们用它来给照片分类、识别未知花卉，或辅助驾驶汽车。但这些强大的系统并不总是以我们（人类）的方式“看”世界，它们有时会以令人惊讶的方式行事。

例如，一个能识别数百个汽车制造商和型号的人工智能系统，可能仍然无法捕捉到汽车和飞机之间的“共性”——即，它们都是主要由金属制成的大型交通工具。

为了更好地理解这些差异，我们今天在《自然》（Nature）杂志上发表了一篇新论文，分析了人工智能系统在组织视觉世界方面与人类的重要区别。我们提出了一种方法，能将这些系统更好地与人类知识对齐（aligning），并表明解决这些差异可以提高它们的鲁棒性（robustness）和泛化能力（ability to generalize）。这项工作朝着构建更直观、更值得信赖的人工智能系统迈出的一步。

【解读】 这篇报告的开篇就直指一个当前AI领域最核心的问题：“**对齐**”（**Alignment**）。但请注意，这里谈论的“对齐”与我们常在新闻中听到的“AI伦理对齐”（即确保AI的价值观和目标与人类一致）不同。本文探讨的是一个更根本的**“认知对齐”**（Cognitive Alignment）——即AI“看见”和“理解”世界的基本方式，是否与人类的认知结构一致。

作者们开宗明义地指出，AI的视觉系统虽然在特定任务上（如识别花卉、车牌）很强大，但它们的“世界观”与人类有本质区别。这导致了AI的“脆弱性”：它们“有用”，但不“鲁棒”和“可靠”。



“鲁棒性”（Robustness）和“可靠性”（Reliability）是AI工程中的“圣杯”。“鲁棒性”指的是AI在面对“没见过”的新情况或轻微干扰时，性能是否会急剧下降。一个“不鲁棒”的AI，可能在训练时能99%准确识别“汽车”，但如果照片是黑白的、下雨的、或者汽车是卡通画风的，它就可能彻底“失明”。

本文的核心论点是：AI之所以“不鲁棒”，是因为它们的“视觉表征”出了问题。

**** “表征” (Representation) ****是理解这篇文章的钥匙，它指的是AI的“内心世界”或“思维模型”。当你（人类）看到“汽车”和“飞机”时，你脑中会立刻激活一个更高层级的抽象概念——“交通工具”或“金属制品”。但AI的“表征”是“扁平的”，它只知道“汽车”（标签A）和“飞机”（标签B）是两个不相关的点，它无法理解这两个点都属于“交通工具”这个“大陆板块”。

因此，本文的目标就是“重组”AI的“内心地图”，让它从一堆混乱、扁平的点，变得像人类一样具有层次化、结构化的“知识大陆”。研究者认为，只有AI的“世界观”与人类“对齐”了，它才能在面对新事物时，做出“直观”且“值得信赖”的判断。



【译文】为什么AI在“判别‘异类’”任务上举步维艰

当你看到一只猫时，你的大脑会创建一个“心理表征” (mental representation)，捕捉关于这只猫的一切——从它的颜色、毛茸茸的皮毛等基本概念，到它的“猫性” (cat-ness) 等高级概念。

人工智能视觉模型也会产生“表征”，它们通过将图像“映射” (mapping) 到一个“高维空间” (high-dimensional space) 中的“点” (points) 来实现这一点。在这个空间中，相似的物品（如两只羊）被放置在相近的位置，而不同的物品（如一只羊和一块蛋糕）则相距甚远。

为了理解人类和模型的“表征”在组织方式上的差异，我们使用了认知科学中经典的“判别‘异类’”任务 ("odd-one-out" task)。我们要求人类和模型从给定的三张图片中，选出“不合群”的那一张。这个测试揭示了它们（人类或模型）“认为”哪两项最相似。

【解读】 这里，作者深入阐释了“表征”在人类大脑和AI模型中的具体差异，并引入了他们的核心诊断工具：“判别‘异类’”任务。

首先，我们来拆解“高维空间中的点”这个比喻。这是一个非常重要且形象的描述。想象一张二维的平面地图，我们可以把“北京”和“上海”在地图上标为两个点，这两个点在地图上的“距离”，反映了它们在真实世界中的“地理关系”。AI的“表征空间”也是一张地图，但它不是二维的，而是可能有数千、数万维（“高

维”)。当AI“看到”一张羊的图片时，它会将其转换为这个高维地图上的一个“数学坐标点”。

AI“理解”世界的方式，就是通过组织这张地图。在理想情况下，所有“羊”的图片，无论什么姿势和品种，都应该在这张地图上聚集在一个“羊的区域”；所有“蛋糕”的图片都应该聚集在“蛋糕的区域”；并且“羊区域”和“蛋糕区域”应该相距很远。

而“判别‘异类’”任务 (Odd-One-Out Task)，就是对AI这张“内心地图”的“CT扫描”。当给AI三张图 (如：羊A、羊B、蛋糕A) 时，AI会计算它们在地图上的“坐标点”。然后，它会看哪两个点“距离”最近。在这个例子中，“羊A”和“羊B”的点应该紧挨着，而“蛋糕A”的点则很远。因此，AI会报告：“蛋糕A是异类”。

这听起来很完美，但问题出在，AI组织这张地图的“依据”是什么？人类是依据“猫性” (cat-ness) 这样的“高级抽象概念”来组织的。而AI的依据，正如我们接下来会看到的，往往是“低级、表面”的特征。这个“判别‘异类’”测试，就一面“照妖镜”，它能精准地“照出”AI的认知结构是“深邃的”还是“肤浅的”。

【译文】 有时候，所有人都同意 (标准答案)。比如给定一张“獭” (tapir)、一只“羊”和一块“生日蛋糕”，人类和模型都能可靠地选出“蛋糕”是异类。

其他时候，正确答案并不明确，人类和模型的看法也不一致。

有趣的是，我们还发现了许多“人类强烈同意一个答案”，但“AI模型却答错了”的情况。

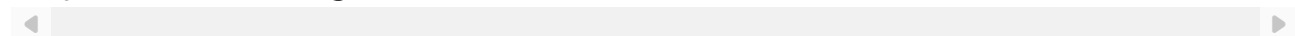
例如在下面的第三个例子中，大多数人同意“海星”是异类。但是，大多数视觉模型更关注“背景颜色”和“纹理”等“表面特征” (superficial features)，因此它们反而会选择“猫”。

(图示) “判别‘异类’”任务的三个例子。三行中分别展示了三张自然界主体的图像。

案例类型	图片1	图片2	图片3	人类选择	模型选择
一致 (Aligned)	獭	羊	蛋糕	蛋糕	蛋糕

案例类型	图片1	图片2	图片3	人类选择	模型选择
不明确 (Unclear)	(图A)	(图B)	(图C)	不一致	不一致
不一致 (Unaligned)	海星	猫	(某海洋哺乳动物，如海狮)	海星	猫

【解读】这是本研究的“第一个关键证据”：作者们找到了一个“确凿的证据” (smoking gun)，证明了AI的“思维方式”与人类存在“系统性错位” (systematic misalignment)。



这个“海星 vs. 猫”的案例极具启发性。为什么人类会“强烈同意”海星是异类？因为我们立刻进行了“抽象分类”：猫和海狮（或其他海洋哺乳动物）都属于“哺乳动物”这个大类，它们都有皮毛（或类似的皮肤质感），是温血的（我们脑补的知识）。而“海星”是“棘皮动物”，从生物学分类上就与前两者截然不同。这是一个基于“语义” (Semantic) 和“概念” (Conceptual) 的判断。

那AI为什么会选“猫”呢？文章一针见血地指出：AI更关注“表面特征”

(Superficial Features)，如“背景颜色”和“纹理”。我们可以合理推测：那张“海星”和“海狮”的图片，很可能都是在“海滩”或“岩石”上拍摄的，它们共享了相似的“背景颜色”（如褐色、蓝色）和“纹理”（如岩石、水面）。而那只“猫”，可能是在“家里的地毯”上拍摄的。



AI并没有“理解”图片中的“主体”是什么。它只是做了一个“统计关联”：图片A和图片C的“背景像素”很像，而图片B的“背景像素”完全不同，所以B（猫）是异类。

这就是AI的“阿喀琉斯之踵”——“捷径学习” (Shortcut Learning)。AI并没有学会“什么是猫”，它只是学会了“互联网上被标记为‘猫’的图片，通常具有哪些统计特征（比如，经常和‘地毯’一起出现）”。这种“捷径”在训练数据上可能很管用，但一旦到了现实世界（比如一只在沙滩上的猫），AI就会彻底失效。这证明了AI的“表征”是肤浅的，它缺乏人类那种“高级抽象”的能力。



【译文】 这个例子说明了人类和AI之间的“系统性错位” (systematic misalignment)，我们在许多不同的视觉模型中都观察到了这一现象，从图像分类器到无监督模型。

这个普遍问题可以从AI“内部地图”的“二维投影” (two-dimensional projection, Pca) 中看出来。

如下左图所示，我们展示了一个视觉模型的“内部地图”，它看起来“毫无结构” (unstructured)，不同类别（如动物、食物和家具）的“表征”都混杂在一起。

右边的结构是我们应用了“对齐方法”后“改进的表征地图”，其中各个类别被清晰地组织了起来。

(图示) 两张地图展示了一个视觉模型对许多不同类别物体的表征。

- **左图（未对齐的分类器 - Unaligned classifier）：**（一个由各种颜色的点组成的“混乱星团”，点（代表动物、食物、家具等）完全混合在一起，没有可见的组织结构。）
- **右图（对齐后的分类器 - Aligned classifier）：**（一个“结构化星图”，点根据类别（如“食物”、“水果”、“家具”、“机械/车辆”）清晰地聚集成成了不同的“大陆板块”，组织有序。）

【解读】 如果说上一个“海星”案例是“症状”，那么这里的“内部地图”就是“病理切片”，它让我们直观地看到了AI“大脑”中的“病灶”。

首先，我们需要理解什么是“二维投影 (PCA)”。AI的“表征空间”（它的“内心地图”）是“高维”的（比如1024维），人类无法直接用肉眼观察。PCA（主成分分析）是一种数学“降维”技术，它就像是把一个立体的“地球仪”（高维空间）“压扁”成一张“世界地图”（二维平面）。这个过程虽然会有“扭曲”（就像地图边缘的格陵兰岛被拉得很大），但它能让我们“一窥”高维空间中的“核心结构”。

****左图（未对齐）****就是AI“病了”的证据。这张“地图”“毫无结构”、“一片混乱” (unstructured jumble)。这意味着，在AI的“世界观”里，“动物”、“食物”和“家具”的“表征”（那些数学坐标点）是“随机”混杂在一起的。在它的地图上，“狗”这个点可能离“热狗”这个点很近，而离“狼”这个点却很远。这**视觉上**证实了AI缺乏我们人类拥有的“高级概念结构”——它没有“动物大陆”、“食物大陆”和“人造物大陆”。它只有一锅“概念汤”。

****右图（对齐后）****则是“治愈后”的理想状态。这正是本文研究要达到的目标。在这张地图上，AI的“内心世界”被“重组”了。所有的“食物”和“水果”聚集在地图的一个角落，形成了“美食大陆”；所有的“车辆”和“机械”聚集在另一个角落，形成了“工具大陆”。

这组对比图的“冲击力”极强。它告诉我们，当前AI的“强大”只是一种“假象”——它是一个“博闻强识的傻瓜”，记住了海量“事实点”，却没能将这些点“结构化”为“知识地图”。而本文的目标，就是给AI装上一个“罗盘”，帮助它“绘制”出这张具有“人类常识结构”的地图。



【译文】 一个多步骤的对齐方法

认知科学家们已经收集了“THINGS”数据集，其中包含了数百万个人类的“判别‘异类’”判断，我们本可以利用它来帮助解决视觉对齐问题。不幸的是，这个数据集“只使用了几千张图片”——这点信息“不足以”直接“微调”（fine tune）强大的视觉模型，（如果强行微调）模型会“立即”在这些小数据集上“过拟合”（overfit），并“忘记”（forget）它们之前的许多技能。

为了解决这个问题，我们提出了一个“三步法”：

1. 我们从一个强大的“预训练视觉模型”（SigLIP-SO400m）开始，并使用“THINGS”数据集在其之上“仔细训练”了一个“小型适配器”（small adapter）。通过“冻结”（freezing）主模型并仔细“正则化”（regularizing）适配器训练，我们创建了一个“教师模型”（teacher model），它“不会忘记”其先前的训练。
2. 这个“教师模型”随后充当“类人判断的代言人”（stand-in for human-like judgments）。我们用它来生成一个名为“ALIGNET”的“海量新数据集”，该数据集使用一百万张不同的图像，包含了数百万个“类人”的“判别‘异类’”决策——这远超我们能从真人那里收集到的数量。
3. 最后，我们使用这个新数据集来“微调”其他AI模型（即“学生”）。由于我们数据集的“多样性”，“过拟合不再是问题”，“学生”可以得到“充分训练”，并能更“深刻地重塑”（deeply restructure）它们的“内部地图”。

（图示） 我们的三步模型对齐方法示意图。

- **步骤1：使用THINGS进行线性对齐。**

- （一个大型视觉模型 + THINGS数据集 → 训练一个小型适配器 → 得到“教师模型”）
- **步骤2：生成大型合成数据集（ALIGNET）。**
 - （“教师模型” + 海量图片 → 生成“ALIGNET”数据集）
- **步骤3：微调“学生”模型。**
 - （“学生”模型的“混乱地图” + ALIGNET数据集 → 充分训练 → “学生”模型拥有了“层次化结构”的新地图）

【解读】这是本文的“核心创新”，即作者团队设计的“解决方案”。他们面临一个经典的“AI困境”：

1. **困境A（数据稀缺）：**高质量的“人类认知数据”（THINGS数据集）非常昂贵且稀少，只有几千张图片。
2. **困境B（灾难性遗忘）：**如果用这几千张图去“强行训练”（微调）一个“知识渊博”的大模型（如SigLIP），大模型会“钻牛角尖”，只记住这几千张图的“答案”，导致“过拟合”（Overfitting）。更糟的是，它会“忘记”它之前从数十亿图片中学到的所有“通用视觉知识”，这在AI领域被称为“灾难性遗忘”（Catastrophic Forgetting）。

作者的“三步法”是一个极其聪明的“知识工程”方案，我们可以用一个“**导师-教材-学生**”的比喻来理解：

步骤1：培养“导师”（教师模型） 你不能直接用“高中生习题”（THINGS数据集）去训练一个“大学教授”（SigLIP大模型），这会导致教授“自废武功”（灾难性遗忘）。作者的妙招是“冻结”（Freezing）教授的“核心知识”（主模型），只让他“戴上一顶新帽子”——那个“小型适配器”（Adapter）。他们只训练这顶“帽子”，让它去学习高中习题（THINGS）中的“人类直觉”。结果就是：你得到了一个“导师”——他既保留了“教授”的所有高深知识（因为主体被冻结了），又新学会了“如何像人一样思考”（适配器）。

步骤2：导师编写“海量教材”（ALIGNET） 现在这个“导师”成了“人类判断的代言人”。我们让他“加班工作”：给他一百万张新图片（远超原来那几千张），让他对“每一组三张图片”都做出“类人”的“异类判断”。于是，这个“导师”创造出了一本“海量的、充满人类智慧的完美教材”（ALIGNET数据集）。这是一种“合成数据生成”（Synthetic Data Generation）技术，它规模化了稀缺的人类智慧。



步骤3：用“完美教材”培养“学生” 最后，我们找一个“新学生”（另一个AI模型），把这本“完美教材”（ALIGNET）交给他。因为这本教材“极其庞大且多样化”，学生无法通过“死记硬背”来“过拟合”。他唯一的出路，就是真正领悟教材中蕴含的“深层规律”——即“人类的层次化概念结构”。结果是：这个“学生”的“内部地图”（表征）被“深刻地重塑”了，从“混乱星团”变成了“有序大陆”（如右图所示）。



【译文】 人类的知识是根据不同“相似性层次”（levels of similarity）来组织的。当我们将模型与人类知识对齐时，模型的“表征”会根据这些“相似性层次”发生改变。这种“重组”（reorganization）遵循了认知科学中已知的人类知识的“层次化结构”（hierarchical structure）。

在对齐过程中，我们看到“表征”会根据它们在人类“概念层次”（conceptual hierarchy）中的“概念距离”（conceptual distance）而“相互靠近”或“相互远离”。

例如，“两只狗”（同属一个“子类别”）的（表征）会“靠得更近”（距离减少），而一个“猫头鹰”和一个“卡车”（分属不同“超级类别”）则会“离得更远”（距离增加）。

（图表） 如下的线图显示了人类与AI表征之间“相对距离的变化”。非常相似的类别的表征倾向于“靠得更近”，而相似性较低的物体对的表征则倾向于“离得更远”。

表1：根据概念层次划分的表征相对距离变化

概念关系（X轴标签）	示例	相对距离的变化（Y轴）	解释（地图上的变化）
同一子类别 (Same subordinate)	两只不同的狗（如：德牧 vs 贵宾）	大幅减少	在“狗”区域内，点靠得更近
同一基本类别 (Same basic)	一只狗和一只猫	小幅减少	“狗”区域和“猫”区域相互靠近
不同基本类别 (Different basic)	一只狗和一只鸟	小幅增加	“狗”区域和“鸟”区域轻微远离

概念关系（X轴标签）	示例	相对距离的变化（Y轴）	解释（地图上的变化）
不同超级类别 (Different superordinate)	一只猫头鹰和一辆卡车	大幅增加	“动物大陆”和“车辆大陆”被远远推开

【解读】这是“治愈”有效的**第一个“客观证据”**。上一个“三步法”是“治疗方案”，而这张表（源自原文的线图）就是“疗效CT片”。它证明了“学生”AI的“内心地图”确实被“深刻重塑”了。

这张表揭示了一个“涌现”（Emergence）现象。研究者在“步骤3”中训练“学生”时，并没有明确“告诉”它：“狗和猫属于‘基本类别’，猫头鹰和卡车属于‘超级类别’”。AI的训练任务仅仅是“判别异类”而已。

然而，为了“最高效”地完成数百万次“类人判断”，AI“学生”自行领悟到了：
“哦，原来人类的世界观是‘层次化’的！”

它“主动”地在自己的“高维地图”上进行了“地理大重组”：

1. 它把“德牧”和“贵宾”的坐标点“拉近”（相对距离大幅减少），因为这能帮它更快地判断“德牧、贵宾、卡车”中的“卡车”是异类。
2. 它把“动物大陆”（猫头鹰）和“车辆大陆”（卡车）的坐标点“推远”（相对距离大幅增加），因为这能帮它更快地判断“猫头鹰、乌鸦、卡车”中的“卡车”是异类。

这个“重组”过程是“自发”的，它遵循了人类的“概念层次”（conceptual hierarchy）。**这太重要了！**这意味着AI不仅仅是在“模仿”（mimic）人类的“答案”，它是在“学习”（learn）人类认知结构中的“深层规则”。它从“死记硬背”的“捷径学习者”，转变成了“理解结构”的“深度学习者”。它那张“混乱”的“概念汤”（左图）终于演变成了“井然有序”的“知识大陆”（右图）。

【译文】测试我们对齐后的模型

我们在许多“认知科学任务”上测试了我们“对齐后”的模型——包括像“多重排列”（multi-arrangement，通过相似性排列许多图像）这样的任务，以及一个我们

自己收集的、名为“层次”（LEVELS）的“新‘判别异类’”数据集。在每一种情况下，我们对齐后的模型在“人类对齐”方面都表现出“显著的提升”，在各种视觉任务中与人类判断的一致性“大幅提高”。

我们的模型甚至学会了一种“类人的不确定性”（'Human-like' uncertainty）。在测试中，“模型的决策不确定性”与“人类做出选择所需的时间”表现出“强相关性”——（人类反应时间）是（衡量人类）不确定性的一个常用替代指标。

【解读】这是“治愈”有效的**第二个“惊喜证据”**——“类人的不确定性”。这在AI安全和可信赖领域是一个“极其重要”的发现。

目前AI最大的危险之一，就是它们“非常自信地犯错”（confidently wrong）。一个“未对齐”的AI，在面对一个模棱两可、极其困难的分类任务时，它仍然会“毫不犹豫”地给出一个“99.9%”的错误答案。因为它在训练时，被告知“必须给出一个高分答案”。

而这个“对齐后”的“学生”模型，则表现出了更高级的“认知智能”。研究者发现：

1. **对于“简单”任务**（如：羊、羊、蛋糕）：人类会“立刻”做出反应（反应时间极短）。此时，“对齐的AI”也会“非常自信”地（低不确定性）给出正确答案。
2. **对于“困难”任务**（如：獾、羊、猪？——三者都是哺乳动物，很难选）：人类会“犹豫不决”（反应时间很长）。此时，“未对齐”的AI可能会瞎猜一个并“假装自信”，而这个“对齐的AI”则表现出了“高不确定性”。

AI的“不确定性”与人类的“犹豫时间”产生了“强相关性”。这意味着AI“知道了“什么是难题”。它学会了“**知道自己不知道什么**”（Knowing what it doesn't know）。

这在AI领域接近于一种“**元认知**”（Metacognition）能力。这对构建“值得信赖的AI”至关重要。我们不需要一个“什么都懂、从不犯错”的AI（这不可能）。我们需要一个“知道自己认知边界”的AI。一个在关键时刻（比如在自动驾驶中遇到它无法识别的障碍物时）会“感到不确定”，并“请求人类帮助”的AI，远比一个“自信地撞上去”的AI要安全得多。本文的研究表明，“认知对齐”是实现这种“安全不确定性”的有效途径。



【译文】我们还发现，让模型“更像人类” (human-aligned)，也会使它们在“整体”上成为“更好的视觉模型”。

我们对齐后的模型在各种“挑战性任务”上表现得“好得多”，例如从单个图像中学习一个新类别（“少样本学习”，"few-shot learning"），或者即使在被测试的图像类型发生变化时也能做出可靠的决策（即“分布外”，"distribution shift"）。

（图表） 两个柱状图显示，我们对齐后的模型（深蓝色）在“认知科学任务”（顶部）和“AI泛化任务”（底部）上的表现均优于原始模型（浅灰色）。

表2：认知科学对齐任务表现

任务（人类对齐度）	原始模型（准确率）	对齐模型（准确率）	表现提升
三元组判别异类	≈ 0.84 (84%)	≈ 0.90 (90%)	+7%
多重排列	≈ 0.28 (28%)	≈ 0.52 (52%)	+86%

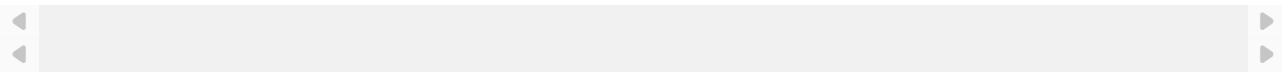
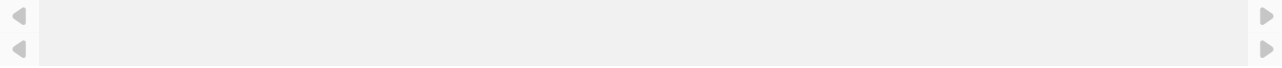


表3：AI泛化能力任务表现

任务（泛化能力）	原始模型（准确率）	对齐模型（Alignet）（准确率）	表现提升
少样本学习	≈ 0.27 (27%)	≈ 0.38 (38%)	+41%
分布外（鲁棒性）	≈ 0.04 (4%)	≈ 0.08 (8%)	+100%



【解读】 这是本文的最终“论点闭环”，也是最关键的“工程学意义”所在。作者证明了：追求“像人一样思考”（认知对齐，表2）并不是一个“不切实际”的“哲学爱好”，它反而是解决AI“核心工程难题”（泛化能力，表3）的“最佳路径”。

表2（认知任务）首先证明“治疗方案”是有效的。在“多重排列”这个更复杂的认知任务上，“对齐模型”的性能暴涨了86%，证明它确实在“认知结构”上更像人了。

表3（泛化任务）则回答了“所以呢？（So What?）”的问题。“像人一样思考”带来了两个“立竿见影”的好处：

1. **更强的“少样本学习”能力 (+41%)**：“少样本学习”（Few-shot Learning）是AI的“软肋”，却是人类的“强项”。一个（人类）小孩只需看过一次“长颈鹿”，就能在动物园里认出它。而传统AI则需要“数千张”照片。为什么“对齐模型”学得更快了？因为它不再是“一张白纸”。它已经拥有了一张“结构化”的“内心地图”（在“步骤3”中学会的）。当你给它看一张“长颈鹿”的照片时，它能立刻“定位”：哦，这个新事物属于“动物大陆”，在“哺乳动物”区域，它有“四条腿”和“长脖子”。这个“预先存在的概念框架”让它能“举一反三”。而“原始模型”的“混乱地图”则没有这种框架，新知识对它而言只是另一个“孤立的数据点”。
2. **更强的“分布外”鲁棒性 (+100%)**：“分布外”（Distribution Shift）是AI“鲁棒性”的“终极考验”。“分布内”指的是“训练数据”（如：高清照片里的猫）。“分布外”指的是“没见过的新类型数据”（如：素描画的猫、毕加索画的猫）。“原始模型”在这项上的得分只有“4%”，几乎等于“彻底失败”。为什么？因为它依赖“捷径学习”（见“海星”案例），它只学会了“高清照片里猫的纹理”。当它看到一张“素描”时（没有纹理），它的“捷径”失效了。而“对齐模型”的性能“翻了一番”(+100%)。虽然0.08的绝对分仍然很低，但这个“相对提升”是巨大的。这证明“对齐模型”在一定程度上“被迫”放弃了“表面纹理”的捷径，转而学习“更深层”的“猫的结构和形状”。这种“更接近本质”的理解，自然比“表面捷径”要“鲁棒”得多。



【译文】 迈向更人性化、更可靠的模型

许多现有的视觉模型“未能捕捉”到人类知识的“高级结构”（higher-level structure）。

本研究为解决这一问题提供了一种“可能的方法”，并表明，模型可以“更好地与人类判断对齐”，并在各种“标准的AI任务上”表现得“更可靠”。

虽然还有更多的对齐工作有待完成，但我们的工作展示了“朝着更鲁棒、更可靠的AI系统迈出的一步”。

【解读】 文章的结论温和而有力。它为我们总结了这场“AI认知重组手术”的完整旅程和深远意义：

1. **诊断“病症”**：我们首先通过“判别异类”任务（海星 vs 猫）和“脑部CT”（PCA降维地图）发现，现有的AI视觉模型虽然强大，但其“内心世界”是“混乱、扁平、无结构”的（a jumbled mess）。它们依赖“表面特征”走“捷径学习”的“邪路”，而不是像人类一样理解“高级抽象结构”。
2. **研发“药物”**：面对“高质量人类数据稀缺”和“灾难性遗忘”的“并发症”，研究者设计了“三步法”这一“特效药”。这个“导师-教材-学生”的知识传递管线，巧妙地将“稀缺的人类智慧”“规模化”为“海量合成教材”（ALIGNET），从而“迫使”“学生”AI去学习“如何像人一样思考”。
3. **验证“疗效”**：
 - **疗效一（结构重组）**：AI的“内心地图”被“深刻重塑”。它“自发”地学会了人类知识的“层次化结构”（如：“动物”和“车辆”是不同“大陆”），如表1所示。
 - **疗效二（认知深化）**：AI学会了“类人的不确定性”，它开始“知道自己不知道什么”，这是迈向“可信赖AI”的关键一步。
 - **疗效三（工程突破）**：“像人一样思考”的AI，在“少样本学习”（学得更快）和“分布外鲁棒性”（更抗干扰）等“核心工程难题”上，表现“好得多”，如表3所示。

最终启示：这篇报告的“震撼”之处在于，它用“数据”和“实验”证明了一个深刻的假说：**通往“强大、鲁棒、可靠”的“人工智能”（Artificial Intelligence）的道路，或许必须经过“认知对齐”（Cognitive Alignment）这座桥。**我们过去试图构建一种“外星智能”（Alien Intelligence），希望它能“另辟蹊径”解决问题。但这项研究表明，至少在“视觉”这个领域，最“强大”的智能结构，可能就是“人类”的智能结构。我们想要AI变得“更可靠”，也许首先得让它变得“更像我们”。



This content was created by another person. It may be inaccurate or unsafe. [Report unsafe content](#)
