假设检验与机器学习：使用 Cohen 的f^2解读深度人工神经网络中的变量效应

Wolfgang Messner 达拉·摩尔商学院，南卡罗来纳大学，哥伦比亚市——美国南卡罗来纳州

深度人工神经网络在许多领域都展现出卓越的预测性能，但无法进行统计推断，其黑箱运作也过于复杂，难以为人类所理解。由于在科学实验和研究模型中，提出某种关系存在往往比预测更为重要，因此机器学习的使用频率远低于推断统计学。此外，统计学还要求通过展示所研究现象的大小来改进理论检验。本文扩展了当前的可解释人工智能方法，并为机器学习开发了一种模型无关的假设检验框架。首先，调整 Fisher变量置换算法，以计算与普通最小二乘回归模型中 Cohen 的 f_2等效的效应量指标。其次，将 Mann-Kendall单调性检验和 Theil-Sen估计量应用于 Apley的累积局部效应图，以确定变量影响的方向及其统计显著性。本文通过一个人工数据集和一项社会调查，并借助 Python沙盒实现，展示了该方法的实用性。

关键词：人工神经网络（ANN）；深度学习；效应量；可解释人工智能（XAI）；统计显著性

Wolfgang Messner 是南卡罗来纳大学达拉·摩尔商学院的国际商务临床教授。他在卡塞尔大学（德国）获得经济学与社会科学博士学位，在威尔士大学（英国）获得财务管理工商管理硕士学位，并先后在慕尼黑工业大学（德国）、纽卡斯尔大学（英国）和佩鲁贾外国人大学（意大利）学习，获得计算机科学与工商管理理学硕士和理学学士学位。可通过 wolfgang.messner@moore.sc.edu 联系他。

1. 引言

自英国统计学家弗朗西斯·高尔顿爵士于1885年将相关和回归定义为统计学概念，卡尔·皮尔逊随后开发出相关性测量指标以来，1895 [1], 各学科的大多数研究方法都旨在确立加性充分性，即找出平均而言对结果具有线性且显著贡献的变量 [2]。这些概念通过 p 值衡量效应的统计显著性，也就是效应是否存在[3]。在这种经典的零假设显著性检验统计方法中，评估的重点更多在于数据与模型的拟合程度，而非利用变量预测结果。但期刊编辑日益呼吁研究人员评估所研究现象的大小，而不仅仅关注其统计显著性 [4] ，以便构建“更贴近从业者需求且更有用”的“基于研究的知识”[5]。Cohen 的f^2 [3,6]是一项极具参考价值的指标，既可评估模型的总体效应量，也可评估各变量的局部效应量，即在整个模型的背景下，各变量实质显著性的效应量。然而，现有研究文献中效应量指标的报告仍不一致[7]。

尽管大多数有趣的研究问题都涉及变量之间的统计关系，但许多复杂过程所依据的数据集呈现出高度非线性行为。然而，普通最小二乘回归分析（OLS）将视角局限于“等距量表上的直线关系，并假定观测值服从正态分布”[3]。这一局限促使研究人员尝试机器学习方法，例如深度人工神经网络，它们能够

近似任意复杂的数学结构，同时对噪声和故障也更具容忍度[8]。过去几年间，各种机器学习库相继开发出来，使非专业人士也能将机器学习应用于各种问题并提取特征[9,10]。然而，机器学习模型在研究和现实决策中的应用仍然有限，因为这些模型将其内在机制隐藏在黑箱中[11,12]。要评估自变量（输入变量）与因变量（输出变量）之间关系的强度、方向和形式并非易事。因此，越来越多关于可解释机器学习（即可解释人工智能（XAI））的研究致力于开发相关技术，使系统变得可解释、透明且易于理解[13–15]。其中包括尝试揭示输入对输出变量的贡献。不过，“目前尚未出现一种实用且适用范围广泛的可解释人工智能技术”[16]。

各学科的研究人员都热衷于比较传统统计推断模型（如 OLS回归）与一般机器学习方法，尤其是神经网络的性能 [例如，17–19]。上述文献在报告其普通最小二乘回归模型时，会说明变量影响的方向和强度、统计显著性以及效应量。然而，使用机器学习模型时，Cohen 的 f^2 遗憾的是无法通过常用的解释工具直接获得。因此，研究人员采用新的可解释人工智能方法，例如变量置换算法，来区分重要变量与不那么重要的变量 [20,21]。但这些变量重要性度量大多是可视化结果，并不等同于推断统计学中使用的统计显著性和实质显著性检验。遗憾的是，缺乏统计上适当的度量，削弱了机器学习和可解释人工智能在科学实验和研究模型中用于假设检验的认可度。

另一类近期提出的可解释人工智能方法包括部分依赖 [22], 局部依赖和累积局部效应图 [两者 23]。这些图以可视化方式展示自变量对模型预测值的影响，从而提高模型可解释性。当自变量之间高度相关时，累积局部效应图比部分依赖图效果更好。此外，它们的计算效率也更高 [24]。但所有这些新颖的可解释人工智能方法主要是为通过可视化支持探索性分析而设计的。尽管“可视化是最强大的解释工具之一” [22], ，全面的统计框架仍应依托量化、可视化和假设检验这三者 [25]。

本文旨在将这些当代可解释人工智能方法与经典统计推断中关于方向、显著性和强度的度量联系起来。基于 Fisher变量置换算法，我为研究人员提供了一种简便的方法，用于估计深度学习网络输入变量的 Cohen效应量 f^2 。此外，我概述了如何将 Mann-Kendall单调性检验 [26,27] 应用于 Apley的累积局部效应图，从而确定输入变量的影响方向和统计显著性。这样一来，我便能在经典推断统计学的完整意义上，使用机器学习进行假设检验。我使用深度学习开源软件库 Keras 2.4.3 和 dalex模型无关解释器 [24,28], ，提供了一个 Python 沙盒实现，供其他研究人员熟悉这一方法。我希望为研究人员提供一种途径，以深度学习人工神经网络取代 OLS回归方法，从而分析科学实验和研究模型中复杂且可能呈非线性的回归问题。

本文其余部分安排如下。下一节将回顾推断统计学与机器学习之间的差异，以及可解释人工智能在使算法模型具备可解释性方面所发挥的作用。随后，我将提出一个机器学习中的假设检验框架，并给出实用的解释指南。在实例一节中，我将介绍该框架在一个人工数据集和一个真实世界数据集上的两项应用。之后将讨论研究结果，并提出未来研究的建议。

2. 推断统计学与机器学习

尽管所有统计学都始于数据，但从数据得出结论有两种截然不同的方法[29]。第一种是推断统计学，它假定数据由一个随机数据模型生成。模型参数根据数据估计得出。随后利用这些参数的方向和显著性水平，通过假设检验对模型进行验证。当多变量 OLS 线性回归、逻辑回归或 Cox比例风险回归“拟合数据以得出定量结论时，结论针对的是模型的机制，而非自然界的机制。因此，如果模型对自然界的模拟效果不佳，结论就可能是错误的”[29]。

第二种方法是算法建模，也称为机器学习。人工神经网络是机器学习的一种方法，能够在没有关于数据中潜在关系性质的先验知识的情况下识别模式 [30]。训练算法通过输入人工神经网络的数据进行学习；人工神经网络是一种灵活、动态的非线性黑箱模型 [31]。这种方法不假设数据来自多变量分布。传统机器学习在处理原始输入数据方面的能力仍有一定局限，而深度学习这一新领域则允许设计由多个处理层构成的神经网络模型，这些处理层能够学习具有多个抽象层次的数据表征 [32]。计算科学家于2006年成功训练出首批实验性深度神经网络，随后在2010 [33–36]之后出现了具有商业可行性的应用。总而言之，机器学习研究将重点从数据模型转向算法的属性，这些属性以其收敛性（在迭代训练周期中）和预测准确率（针对新的、迄今未见过的数据）为特征。不过，“尽管预测结果令人信服，但缺乏明确模型可能使机器学习解决方案难以直接与现有知识联系起来” [37]。

可解释人工智能（XAI）旨在通过开发模型解释与阐释技术来应对机器学习的这一缺陷 [13–15]。与推断统计学中针对变量的显著性检验和效应量检验类似，模型无关解释器试图描述不同解释变量对结果的贡献。早期的 XAI 度量方法旨在揭示输入变量对输出变量的影响 [38], , 例如，总体影响度量 [GIM 会考察网络权重的数值；39], 或者依次将每个输入变量对应的权重归零，以评估神经网络的响应 [SZW；40]。然而，在深度学习中，权重矩阵构成的复杂系统连接着众多隐藏层，因此简单的 GIM 和 SZW 解释方法已不再够用。近年来，XAI 文献提出了其他方法，其中一些列于表1。这些方法最初旨在帮助研究者理解他们创建的黑箱，并支持探索性分析，但仍严重依赖可视化。因而，从最严格的经典推断统计学意义上来说，它们尚不适用于假设检验。

3. 机器学习中的假设检验框架

现在，我将尝试弥合算法建模与推断统计学之间的差距。基于表1所列的可解释人工智能度量方法，我首先概述一种计算 Cohen效应量指标 f^2 的算法。接着，我将介绍 Mann-Kendall 单调性检验和 Theil-Sen估计量，以判断变量影响的方向及其相关的统计显著性。本节最后，我会提供使用这一机器学习中的新假设检验框架的指南。

表1：现有的数据集级XAI方法

XAI方法	描述	优缺点
变量重要性度量 [20]	衡量模型性能 在移除一个或一组输入 变量后发生的变化。如果某个变量 很重要，那么在对这些变量进行置换后， 模型的性能 会变差。输出为平均 RMSE 损失 经过一定次数的置换后。	该方法的输出很容易 理解和解读。但 RMSE 损失取决于置换的随机性以及 置换方式和损失函数的选择 函数。无法对 可获得统计显著性。
部分依赖图 [22]	显示模型预测期望值 如何随某个 输入变量变化。使用所有 其他条件不变（CP） 用于绘制的来自数据集的观测值 部分依赖图。	提供一种简单的方法来概括 ……的影响 变量的可视化。不适用于不稳定的 相关输入变量 [24]。 量化或统计显著性 检验。
局部 累积局部效应剖面 [23]	预测变化的平均值（但不是 预测值本身），并将其累积 到网格上。	计算快速且无偏。在 特征空间中，估计值的 准确度各不相同。解释起来 很困难 高度相关，因为其效应 无法从单个变量的效应中 区间也是不允许的 [41]

此表概述了适用于深度人工神经网络的现有数据集级XAI方法。

3.1 效应量：Cohen 的 f^2 机器学习

对于自变量和因变量均为连续变量的回归问题，Cohen 的 f^2适用于计算模型的总体效应量 [3,6]：

$$f^2 = \frac{R^2}{1 - R^2}$$

在此方程中，R^2 表示相对于不含任何回归变量的模型，所有自变量所解释的方差比例。然而，更有趣的是，Cohen 的f^2 针对某个感兴趣的单个变量 V [3,42,43]的变体：

$$f_V^2 = \frac{R^2 - R_V^2}{1 - R^2}$$

在此方程中，R_V^2 表示相对于不含任何回归变量的模型，除 V 之外的所有其他变量所解释的方差比例。因此，分子 $R^2 - R_V^2$ 反映了由 V 独特解释的方差，即在所有其他变量所解释的方差之外，V 额外解释的部分。在行为科学中，0.02、0.15 和 0.35 被定义为f_V^2 小、中等和大值的阈值；小于 0.02 的值表示微不足道的影响[6]。由于f_V^2 通常无法直接从常用的统计软件中获取，因此需要通过估计两个不同回归模型中所解释的方差来手动计算：一个包含所有变量的模型（R^2），以及一个不包含 V 但包含所有其他变量的模型（R_V^2）。

由于数据初始化具有随机性，且训练周期需要耗费大量资源，因此这种 loco（逐一排除协变量）方法并不适用于机器学习模型。为此，我建议采用 Fisher变量置换算法 [20,21], , 在模型构建完成后消除某个变量的影响。设 X 为整个数据集。通过随机打乱与 V X_V 相关的数据，可以得到一个经过修改的新数据

集 X_V^*，并据此计算预测结果 $\hat{y}$。如果 V 在模型中很重要，那么置换后，预测结果 $\widehat{y_V^*}$ 的准确性应该会降低。更形式化地说：

1. 将原始数据 X 的均方误差损失函数计算为 $MSE(\hat{y}, X, y)$；2. 通过随机置换 X 的第 V 列，创建一个新的 X_V^*；3. 基于修改后的数据 X_V^*计算模型预测 $\widehat{y_V^*}$；4. 计算 $MSE(\widehat{y_V^*}, X_V^*, y)$；5. 重复步骤 2 至 4 数次（B 次），以消除随机性的影响，并返回平均值$\frac{1}{B} MSE(\widehat{y_V^*}, X_V^*, y)$。

上述步骤可在可用的解释器（例如 dalex）中进行配置和执行。由于效应量公式中的 R^2指的是总体，而非样本值 [3], ，因此我使用 X 的所有可用数据，即合并训练数据集和测试数据集。

接下来，我需要处理输出并计算 R^2 和 R_V^2，以便计算 f_V^2。请注意，以下公式仅在数据集 X 已标准化时才有效；标准化是机器学习算法的常见要求：

6. Compute $R^2 = 1 - MSE(\hat{y}, X, y)$ and $R_V^2 = 1 - MSE\left(\widehat{y_V^*}, X_V^*, y\right)$; and
7. 计算 $f_V^2 = \frac{R^2 - R_V^2}{1 - R^2}$。

与 LOCO 方法相比，变量置换算法可能会高估变量的效应量。前者会从模型中移除一个外生解释变量，并假定我们并不了解它；后者则会随机改变该解释变量，从而扰乱模型的解释能力。这会夸大均方误差的计算结果。此外，不同模型对数据重排的平均响应各不相同，因此，上述计算出的效应量既无法在不同模型之间进行比较，也无法与小、中等和大效应的常用基准阈值进行比较。因此，我建议采用一个调整因子，使通过变量置换过程计算出的效应量能够与 LOCO 方法相比较。

首先，我计算基准 $MSE(\widehat{y^*}, X^*, y)$，其中随机重排的不仅是一个变量，而是模型的所有变量。这个基准均方误差很好地表明了在数据中不存在预测信号 [44] 时可能出现的最差损失函数值：

8. 通过随机置换 X 的所有列创建一个新的X^*；9. 基于修改后的数据X^*计算模型预测 $\widehat{y^*}$；10. 计算MSE（$\widehat{y^*}$、X^*、y）；11. 重复步骤 8 到 10 多次（B 次）以减少随机性，并返回平均值$\frac{1}{B} MSE$（$\widehat{y^*}$、X^*、y）。

与置换单个变量类似，我接下来计算基线R_{base}^2，用于计算基线f_{base}^2：

12. Compute $f_{base}^2 = \frac{R^2 - R_{base}^2}{1 - R^2}$ with $R_{base}^2 = 1 - MSE(\widehat{y^*}, X^*, y)$.

接下来，我通过将其与整个模型的总体效应量$f^2 = \frac{R^2}{1 - R^2}$关联起来，对 f_{base}^2 进行调整。因此，我调整f_V^2：

13. 计算调整因子$a = \frac{f^2}{f_{base}^2}$；14. 调整$f_V^2$：

$= a \cdot f_V^2$。

需要明确的是，最后这一步会根据模型对数据重排的平均响应相较于 LOCO 方法的情况，对 f_V^2 进行调整，使 f_V^2 更便于与 Cohen 原始的 f^2进行比较。因此，在解释时，我建议继续使用 Cohen 原始阈值 [6], ，即：微不足道的影响

效应量：$f_V^2 < 0.02$ 为小，$0.02 \le f_V^2 < 0.15$ 为中等，$0.15 \le f_V^2 < 0.35$ 为大，$f_V^2 \ge 0.35$。请注意，与 R^2 不同，f_V^2 可以大于 1。

这种方法完全不依赖特定模型，也就是说，它可以用于不同类型的机器学习模型。虽然均方根误差（RMSE）损失函数在机器学习应用中更为常用，但改用均方误差损失函数后，所得的f_V^2 度量与普通最小二乘回归模型中的f_V^2 更具可比性。不过，这种方法的主要缺点是依赖置换的随机性。因此，在不同的运行中，报告的f_V^2 值也会有所不同。不过，可以通过多轮置换（即多次重复步骤 2-4 和 8-10，例如 $B = 50$），并取平均值（见上文步骤 5 和 11），来控制随机性并提高准确性。不言而喻，这会增加算法的计算时间。

3.2 影响方向：Mann-Kendall 单调性检验与 Theil-Sen估计量

上述效应量指标 f_V^2 并未提供关于效应方向的任何信息。然而，很自然会想知道因变量的取值平均而言是在上升、下降还是保持不变。此外，我们也很自然会想了解这种趋势是否具有单调性，也就是说，是否贯穿自变量取值的整个范围。为了在假设检验过程中回答这些问题，我建议利用累积局部效应图 [23], ，应用非参数的曼-肯德尔单调性检验 [26,27], ，然后用泰尔-森估计量 [45]量化效应。

Apley的累积局部效应图是一种可视化方法，通过概括自变量对模型预测值的影响，提升监督式机器学习模型的可解释性和透明度。与更流行且历史更久的部分依赖图 [22], 不同，它们无需对其他可能相关的自变量进行不可靠的外推[24]。

Mann-Kendall 检验通过分析与较高自变量和较低自变量对应的因变量之间差值的符号，来评估并验证自变量与因变量之间的联系，由此得到总计 $^{n}_{(n-1)}\quad_{2}$ 对数据，其中 n 是自变量的分辨率（即自变量取值范围内可用观测值的数量）。Mann-Kendall 检验统计~~量适用~~于许多情形，因为它“相对有效且稳健”[46], ，对数据变换（如对数变换）具有不变性，且不要求数据符合任何特定分布[47]。然而，该检验无法抵御序列相关；这可能导致对无趋势零假设的过度拒绝（第一类错误）。因此，我采用滞后阶数为 3 的 Hamed 和 Rao 修正的 Mann-Kendall 检验，通过应用方差校正方法来处理序列相关[48,49]。由于所有校正都会增加第二类错误，Mann-Kendall 检验“最好视为一种探索性方法[.....] [并] 最适合用于识别变化显著或幅度较大的站点，并量化这些发现”[50]。

接下来，我根据累积局部效应图中的预测值，使用 Theil-Sen估计量评估变量的影响方向。该估计量是一种稳健方法，它根据所有数据点对的斜率中位数，确定连接两个变量的线性回归方程的斜率。具体而言，与普通最小二乘线性回归相比，该估计量对离群值、非正态性和异方差性更具稳健性。Theil-Sen估计量通过自助法计算置信区间[45],，我据此计算相应的 p 值[，方法遵循51]所述。

3.3 实践与解释指南

我使用 dalex 作为模型无关解释器，并通过 Keras 和 Python 将这一假设检验框架实现为一个沙盒。我已将其放在在线补充材料中，供其他研究人员

熟悉并用于自己的工作。神经网络模型在数据上训练完成后，dalex 解释器对象会为该模型创建一个包装器。随后，按照上述详细说明的两个步骤对这个包装后的模型进行检查。

针对每个输入变量，计算并列出以下指标：Cohen效应量 f_V^2；单调性检验的 p 值；影响方向的斜率及其 p 值。表2提供了在研究论文中分析和报告这些指标的建议。如果效应并非单调的，继续分析累积局部效应图并进行变点分析，可能会带来更多见解 [52,53]。

表2：假设检验报告指南

单调性显著（p-值为 Mann-肯德尔检验）？	影响的方向（Theil-Sen 斜率和p-值）？	效应量			
		$f_V^2 < 0.02$ (微不足道)	$0.02 \leq f_V^2 < 0.15$ (小)	$0.15 \leq f_V^2 < 0.35$ （中等）	$f_V^2 \geq 0.35$ （大）
$p \leq 0.05$	$s < 0 \wedge p \leq 0.05$	独立的变量仅产生微不足道的影响因变量，该变量呈单调变化，为负向，且在统计上显著。	自变量产生了小的单调效应，作用于因变量，其为负且具有统计显著性。	该自变量产生了中等程度的单调效应对因变量的影响，这种影响为负向且具有统计显著性。	自变量变量产生了大效应，作用于因变量，该效应为负向且在统计上显著。
	$s > 0 \wedge p \leq 0.05$	自变量变量仅产生微不足道的影响因变量，它是单调的，正向的，且具有统计显著性。	自变量变量产生了小且单调的效应作用于因变量，该效应为正向，且具有统计显著性。	自变量变量产生中等效应，作用于因变量变量，它是正向且具有统计显著性。	自变量变量产生大效应，作用于因变量变量，即正向且具有统计显著性。
$p > 0.05$		自变量变量仅产生微不足道的影响于因变量，而因变量并非单调的。	自变量变量产生了小的影响，作用于因变量，该影响并非单调的。	自变量变量产生中等效应，作用于因变量变量，该变量并非单调的。	自变量的作用对……的大效应因变量变量，其不是单调的。

本表提供了使用机器学习假设检验框架的指南。除文字说明外，还应报告以下统计量：Mann-Kendall p 值；效应量 f_V^2；Theil-Sen 斜率 s、p 值。

按照在语境明确时于研究报告中报告特定变量效应量的惯例，我建议省略上标中的 V，即直接使用 f^2，而不是 f_V^2。在下一节的实例中，我将遵循这一惯例。

4. 实例

现在，我将假设检验框架应用于一个人工数据集和一个真实数据集。人工数据集基于一种代数函数形式，即库仑方程。真实数据集来自一项社会科学调查，探讨若干个人因素对主观幸福感的影响。这些数据集在结构上的差异体现在输入变量的数量、函数形式的复杂程度，以及

输入变量之间的相关性。此外，真实数据集还包含程度较高但未知的噪声。

在线补充材料包含数据、用于构建和训练深度网络的 Keras/Python 代码，以及用于使用假设检验框架的 Python 代码。

4.1 第一个示例：库仑定律方程

这第一个数据集展示了一种相当简单的函数形式的行为。库仑平方反比定律，也称为库仑定律方程，描述了两个带电物体之间的静电力 F。静电力的主要特征包括存在两种电荷 q_1 和 q_2，、同种电荷相斥而异种电荷相吸，以及两个物体间距 r 增大时力会减小。其基础数学公式由法国物理学家夏尔·库仑（1736—1806）提出：

$$F = k\,\frac{q_1\,q_2}{r^2} \quad \text{with} \quad k = \frac{1}{4\,\pi\,\varepsilon\,8.854\cdot 10^{-12}}\,.$$

在此公式中，k 是电力（库仑）常数，取决于介质的相对介电常数 ε，例如真空中的值为 1，有机氯化合物二氯甲烷 CH_2Cl_2, 中的值为 9.08，水 H_2O 中的值为 80.1（温度为 20o C 时）。

我创建了一个包含约 125,000 个元组的数据集 ∈ [0；1]。随后，我对数据进行缩放，使得 $q_1, q_2 \in [0；10]$; $r \in [0；1.5]$；以及 $\varepsilon \in [1；80]$。计算得到的输出变量范围为 $F \in [0；1.931]$。由于数据遵循代数函数形式，我可以立即提出四个假设：

H1：电荷 q_1 增加与两个电荷之间静电力 F 增大相关。H2：电荷 q_2 增加与两个电荷之间静电力 F 增大相关。H3：两个电荷之间的间距 r 增大与静电力 F 减小相关。H4：介电常数 ε 增大与两个电荷之间静电力 F 减小相关。

4.1.1 深度网络的计算方面 为检验这些假设，我构建了一个深度人工

神经网络，共有五层：第一层（输入层）有四个节点，第五层（输出层）有一个节点，与数据集相匹配。我对变量进行预处理，使其均值为零、标准差为一。三个隐藏层各有五十个节点。对于输入层和隐藏层，采用线性整流单元（ReLU）作为激活函数，并使用 He正态初始化来初始化权重 [54]。我使用 Glorot均匀初始化方法初始化连接到输出层的权重 [34]。由于这最终是一个标量回归问题，因此我采用均方误差损失函数来匹配网络架构 [55]，并使用高效的 ADAM（自适应矩估计）优化算法 [56]。训练期间，我以 0.001 [57] 的丢弃率随机移除神经网络中的一部分节点。丢弃后，我对剩余节点施加五的权重约束 [58]。使用这些超参数，五十个训练轮次足以训练该神经网络。

4.1.2 该训练数据上的结果

对于经过训练的神经网络，我使用从数据集中随机抽取的 10,000 个元组子集和 B = 50 次置换运行假设检验框架。结果汇总于表3。

完整模型的R^2 为 0.968。电荷q_1 对静电力 F 具有较大的单调效应，该效应为正且具有统计显著性，Mann-Kendall $p < 0.001$；$f^2 = 8.210$；Theil-Senn s =0.007；$p < 0.001$。正如预期，电荷q_2 的效应类似，Mann-Kendall $p < 0.001$；$f^2 = 8.115$；

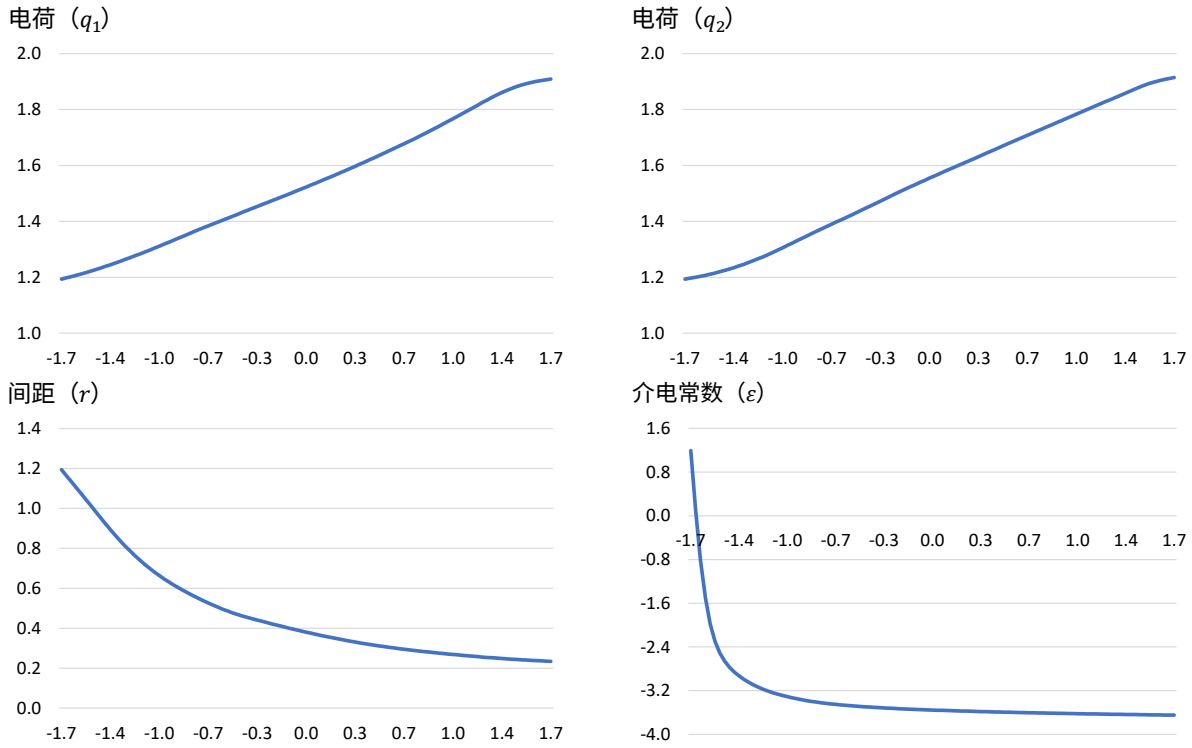
Theil-Senn $s = 0.007$； $p < 0.001$。这支持 H1 和 H2。作为对 H3 的支持，间距 r 对 F 具有较大的单调效应，该效应为负且具有统计显著性，Mann-Kendall $p < 0.001$； f^2 =9.629；Theil-Senn $s = -0.006$； $p < 0.001$。作为对 H4 的支持，介电常数ε 对 F 具有较大的单调效应，该效应为负且具有统计显著性，Mann-Kendall $p < 0.001$； $f^2 = 23.166$；Theil-Senn$s = -0.004$； $p < 0.001$。

表3：变量效应（库仑定律）

变量	f_V^2	单调性 (Mann-Kendall)	*P*	泰尔-森斜率	*p*
电荷 q_1	8.210	上升	< 0.001	7.387 ▪ 10^{-3}	< 0.001
电荷 q_2	8.115	上升	< 0.001	7.772 ▪ 10^{-3}	< 0.001
间距 r	9.629	下降	< 0.001	-6.140 ▪ 10^{-3}	< 0.001
介电常数 ε	23.166	下降	< 0.001	-4.296 ▪ 10^{-3}	< 0.001

本表报告了假设检验框架的结果，示例 1：库仑定律。估计基于 10,00 0 个随机选取的数据 r 记录（约占整个数据集的 8%），采用 B = 50 置换；$R^2 = 0.965$。

图1：累积局部剖面（库仑定律）



这些图展示了示例 1（库仑定律）中所用变量的累积局部剖面。估计基于随机选取的 10,000 条数据记录（约占整个数据集的 8%）进行，并采用 B = 50 次置换。

图1中变量的累积局部剖面可视化提供了进一步的佐证。 q_1 和 q_2 的剖面实际上呈直线状，这表明它们之间存在线性

关系，符合库仑定律。r和ε 的剖面呈单调双曲线，这反映出它们在库仑定律中作为分母。

4.2 第二个示例：主观幸福感

第二个数据集包含在德国开展的一项社会科学调查的 25,700 份答卷，该调查为欧洲社会调查 [ESS；59]。我关注的是个人价值优先事项、观点及其他个体因素如何影响个人幸福感 [60,61]。幸福感是一种情绪性心理状态，由基本心理活动和要素组合而成，并可映射到人脑中的内在神经网络 [62], ，因此，使用深度学习人工神经网络进行表征尤为合适。

不同形式的幸福感在经验上彼此独立 [63], ，因此，界定我在此处所使用的这一概念十分重要。遵循以往关于幸福感的研究 [，例如 64,65] ，以及相关的应用文献 [，例如 66,67], ，我认为主观幸福感包括认知成分和情感成分。认知成分指一个人对生活的总体满意度；情感成分则指其体验积极情绪与消极情绪的频率差异。两者结合起来，通常相当于所谓的日常幸福感与满意度 [65,68]。欧洲社会调查通过两个 11 点李克特量表条目衡量认知和情感幸福感。遵循幸福感研究的指导原则 [67,69], ，我将输出变量主观幸福感计算为这两个条目的简单算术平均值。

作为输入变量，我从 ESS 中选取了 159 个用于描述个体的变量（详见在线补充材料中的变量清单 [70]）。对于示例假设，我选取了四个经典变量；以往的幸福感研究表明，这些变量与主观幸福感相关 [71]。个体之间的收入和教育差异通常与幸福感报告相关，但以往研究表明，这类地位差异对主观幸福感方差的解释比例不超过 10% [72]。ESS 变量 hincfel 询问受访者对当前家庭收入状况的感受（4：非常困难；3：困难；2：勉强应付；1：以目前的收入生活舒适），而 eduyrs 则记录受访者完成的受教育年限。幸福感包含超越经济繁荣的积极因素 [73], ，其中健康或许是影响生活质量的所有维度中最重要的一项 [74]。与生活能力相关的变量通常能够解释约 30% 的共同方差 [72]。因此，health 变量反映的是受访者对自身总体健康状况的主观评价（5：非常差至 1：非常好）。信任是任何社会环境中的重要组成部分，并会影响经济影响、健康和幸福感[，例如，75,76]。变量 ppltrst 询问受访者是否认为大多数人值得信任，还是与人打交道时再谨慎也不为过（0：再谨慎也不为过；10：大多数人值得信任）。

综上，这得出以下假设：H1：较高的收入与较高水平的主观幸福感相关。H2：较高水平的教育与较高水平的主观幸福感相关。H3：较好的总体健康与较高水平的主观幸福感相关。H4：对他人更高的信任与较高水平的主观幸福感相关。

4.2.1 深度网络的计算方面 为检验这些假设，我构建了一个深度网络，

输入层包含 159 个节点，用于描述一名调查受访者；输出层有一个节点，表示此人的幸福感；此外还有四个隐藏层，每层各有 500 个节点。我将分类变量转换为二元占位变量（参见在线补充材料），并将所有其他变量重新缩放到 [0; 1] 的区间。我使用 k近邻法填补少量缺失值（k = 5; 根据距离的倒数对点进行加权）。我对所有输入数据和输出数据进行预处理，使其均值为零、标准差为一。对于输入层和隐藏层，使用 ADAM 优化算法以及采用 He正态初始化的 ReLU激活函数

初始化。对于输出层，则使用线性函数和 Glorot 均匀权重初始化。为了提升模型在小数据集上的泛化能力，我在每一层和偏置项上添加 L_2 正则化项，权重衰减参数设为极小值 $\lambda = 0.001$。训练期间，我以 0.1 的丢弃率随机移除神经网络中的部分节点；丢弃后，我对剩余节点施加权重约束 3。由于所有变量均由李克特量表调查问题的回答得出，我在输入层和输出层添加以零为中心的高斯噪声（标准差为 0.01），作为随机数据增强。这通过体现平滑性假设 [33,78]，有助于提升泛化能力和容错性。训练神经网络只需十个训练轮次。

4.2.2 该训练神经网络上的结果

在训练好的神经网络上，我使用从数据集中随机选取的 10,000 个元组子集和 50 次置换运行假设检验框架。结果汇总于表 4。

完整模型的R^2 为 0.527。变量 hincfel 对幸福感产生微小的单调效应，该效应为负且具有统计显著性，Mann-Kendall $p < 0.001$；$f^2 = 0.116$；Theil-Senn $s = 0.006$；p <0.001。由于 hincfel 采用反向计分，这支持 H1，即较高收入通常与较高水平的主观幸福感相关。支持 H2 的是，接受教育的时间越长（eduyrs），对主观幸福感的影响越负，且该效应是单调的并具有统计显著性，Mann-Kendall p <0.001；$f^2 = 0.004$；Theil-Senn $s = 0.001$；$p < 0.001$。由于效应幅度极小，这种关系微不足道，在实际层面并不相关。“尽管并非毫无关联，但效应的大小和统计显著性在逻辑上是样本数据的彼此独立特征”[3]。变量 health 对幸福感的影响较小，该效应为负且具有统计显著性，Mann-Kendall $p < 0.001$；$f^2 = 0.079$；Theil-Senn $s = 0.005$；$p < 0.001$。由于 health 采用反向计分，这支持 H3。当人们信任他人（ppltrst）时，这会对幸福感产生单调效应；支持 H4 的是，该效应为正且具有统计显著性，Mann-Kendall $p < 0.001$；$f^2 = 0.006$；Theil-Senn $s = 0.136 \cdot 10^{-3}$；$p < 0.001$。与变量 eduyrs 类似，这种效应微不足道，没有实际意义。

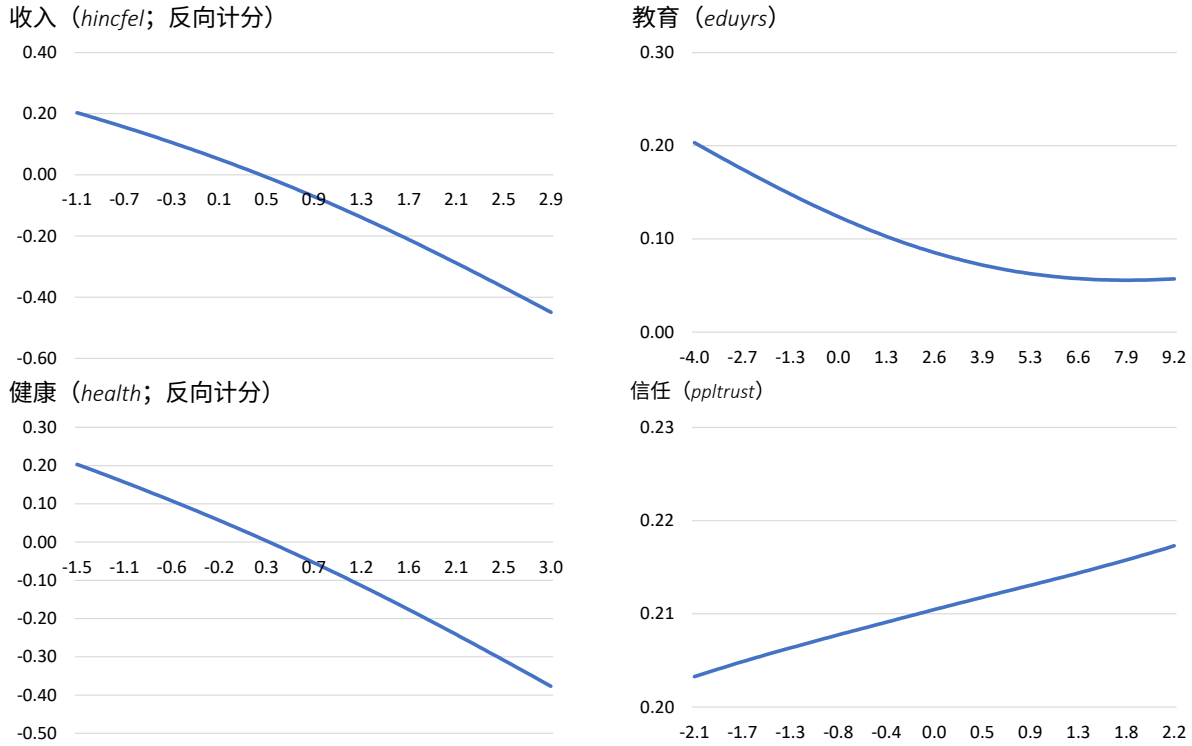
图2中对变量累积局部剖面的可视化提供了进一步的佐证。健康的剖面清楚地呈现出线性关系。hincfel的剖面略微向下弯曲（再次注意，hincfel的计分方向是反向的），这表明高收入者报告的幸福感明显更高。

表 4：变量效应（主观幸福感）

变量	f^2_V	单调性 (Mann-Kendall)	*P*	泰尔-森斜率	*p*
收入（*hincfel*）	0.116	下降	< 0.001	-6.562 ▪ 10^{-3}	< 0.001
教育（*eduyrs*）	0.004	下降	< 0.001	-1.510 ▪ 10^{-3}	< 0.001
健康（健康）	0.079	下降	< 0.001	-5.827 ▪ 10^{-3}	< 0.001
信任（*ppltrust*）	0.006	上升	< 0.001	0.136 ▪ 10^{-3}	< 0.001

本表报告了示例2“主观幸福感”的假设检验框架结果。估计基于随机选取的 10,000 条数据记录（约占整个数据集的 39%）进行，置换次数为 B = 50；$R^2 = 0.527$。

图2：累积局部剖面（主观幸福感）



这些图展示了示例2（主观幸福感）中所用变量的累积局部剖面。估计基于随机选取的 10,000 条数据记录（约占整个数据集的 8%），使用 B = 50 次置换完成。

5. 讨论

研究人员通常不会使用机器学习模型进行假设检验。与多元普通最小二乘回归模型相比，机器学习模型更适合回答许多有关复杂统计关系的有趣研究问题，但它们将统计推理隐藏在黑箱中。可解释人工智能已开发出一些技术，用于评估输入变量和输出变量之间联系的强度、方向和形式。尽管这些新颖的可解释人工智能方法主要借助可视化来区分重要变量与不太重要的变量，但它们与经典推断统计学指标的关联并不理想。事实上，“在机器学习中，可解释性的概念既重要又难以捉摸”[79]。迄今为止，这一问题不幸地削弱了人们在科学实验和研究模型中接受机器学习模型用于假设检验的程度。

本文概述了一种估计深度学习网络输入变量 Cohen’s f^2 的方法，并展示这些变量的影响方向和统计显著性。这使研究人员能够超越多变量 OLS回归，使用并报告他们在推断统计学中熟悉的信息，分析更复杂深度学习网络中的回归问题。在机器学习领域，研究表明，“当解释所需的认知努力较少时，效果会更好” [80] ，这也减轻了研究人员的负担。为支持这一方法，我介绍了一款 Python 软件，它基于深度学习开源软件库 Keras 构建，并扩展了 dalex 的

模型无关解释器接口。

虽然我已在多个数据集上检验了我的假设检验框架，并在本文中记录了两个示例，但仍有更多研究有待开展：

首先，使用变量置换算法估计效应量，并不一定等同于传统的 LOCO 方法。正如我在文章中所解释的，后者并不适用于机器学习。我已提出一些步骤，用于调整前者计算出的效应量。在我测试所用的数据集上，Cohen 提出的效应量阈值似乎是合理的。但显然，这还需要进一步验证。

其次，当机器学习模型的表现非常差时，模型的R^2 可能为负值[81]。这仅表示模型表现不佳，但无法确切知道它的表现究竟有多差，因为R^2 的下限为$-\infty$。显然，在这种模型中对变量效应进行假设检验毫无意义，但我用于计算效应量的方法并未明确处理这种（罕见的）情况。

第三，为了概括自变量对因变量的影响，我对变量的累积局部效应曲线使用了Theil-Sen估计量。我认为，与局部依赖图或部分依赖图相比，累积局部效应曲线更合适，因为它不包含其他自变量的相关效应。我将斜率的置信区间转换为 p 值，以判断斜率是否具有统计显著性。还需要进一步研究和实验来确认这一方法。

第四，在神经网络的语境中，单个自变量很少会独自发挥重要作用 [82]。尽管如此，我尚未考虑一组自变量之间的交互作用。在多元普通最小二乘回归模型中，这往往会导致多重共线性问题。然而，深度人工神经网络冗余的架构能够学习输入变量之间可能存在的相似性，并在多重共线性的情况下进行预测 [83–85]。因此，假设检验框架未来可能的进一步发展应当探讨自变量之间的交互作用。

在线补充材料：代码和数据可用性

假设检验框架、实例和数据集的 Python 代码可在在线存储库 OSF.IO 中获取（数字对象标识符：10.17605/OSF.IO/QDJCY）。

参考文献

[1] J.L. Rodgers，W.A. Nicewander，观察相关系数的十三种方式，《美国统计学家》42（1988）59–66。https://doi.org/10.1080/00031305.1988.10475524。[2] W. Messner，使用激活函数对带噪声的必要条件进行实证评估，《计算机管理科学》18（2021）1–23。https://doi.org/10.1007/s10287-020-00377-2。[3] J. Cohen，P. Cohen，《行为科学中的应用多元回归/相关分析》，第2版，Lawrence Erlbaum，Hillsdale，新泽西州，1983。[4] P.D. Ellis，国际商务中效应量及研究结果的解释，（2010）1581–1588。https://doi.org/10.1057/jibs.2010.39。[5] T.G. Cummings，追求一所投入型学术机构，《管理学会评论》32（2007）355–360。https://doi.org/10.5465/amr.2007.24349184。[6] J. Cohen，《行为科学中的统计功效分析》，第2版，Lawrence Erlbaum，Hillsdale，新泽西州，1988。[7] C.J. Ferguson，效应量入门：临床医生和研究人员指南，《专业心理学：研究与实践》40（2009）532–538。https://doi.org/10.1037/a0015808。[8] M.C. Medeiros，M. McAleer，D. Slottje，V. Ramos，J. Rey-Maquieira，需求估计的替代方法：具有条件波动率的航空旅客到达量高频神经网络回归，《计量经济学杂志》147（2008）372–383。https://doi.org/10.1016/j.jeconom.2008.09.018。[9] K.M. Lee，J. Yoo，S.W. Kim，J.H. Lee，J. Hong，自主机器学习平台，《国际信息管理杂志》49（2019）491–501。https://doi.org/10.1016/j.ijinfomgt.2019.07.003。[10] H. Farbmacher，L. Löw，M. Spindler，用于理赔管理中欺诈检测的可解释注意力网络，《计量经济学杂志》228（2020）244–258。https://doi.org/10.1016/j.jeconom.2020.05.021。[11] Y. Ning，M.E.H. Ong，B. Chakraborty，B.A. Goldstein，D.S.W. Ting，R. Vaughan，N. Liu，用于可解释机器学习的 Shapley变量重要性云，《Patterns》。（2022）100452。https://doi.org/10.1016/j.patter.2022.100452。[12] G. Stiglic，P. Kocbek，N. Fijacko，M. Zitnik，K. Verbert，L. Cilar，医疗保健中基于机器学习的预测模型的可解释性，《Wiley跨学科评论：数据挖掘与知识发现》10（2020）1–13。https://doi.org/10.1002/widm.1379。[13] D. Doran，S. Schulz，T.R. Besold，可解释人工智能究竟意味着什么？视角的新概念化，《CEUR研讨会论文集》2071（2018）1–8。http://ceur-ws.org/Vol-2071/CExAIIA_2017_paper_2.pdf。[14] G. Futia，A. Vetrò，将知识图谱整合到深度学习模型中，以实现更易理解的人工智能——未来研究面临的三项挑战，《Information》11（2020）。https://doi.org/10.3390/info11020122。[15] L. Gianfagna，A. Di Cecco，《使用 Python 实现可解释人工智能》，施普林格，Cham，2021。[16] D. Pedreschi，F. Giannotti，R. Guidotti，A. Monreale，S. Ruggieri，F. Turini，黑箱人工智能决策系统的有意义解释，载于：第33届 AAAI 人工智能大会，2019：第9780–9784页。https://doi.org/10.1609/aaai.v33i01.33019780。[17] G.A. Abbasi，L.Y. Tiew，J. Tang，Y.N. Goh，R. Thurasamy，加密货币作为颠覆性力量的采用：基于深度学习的双阶段结构方程建模与人工神经网络分析，《PLOS ONE》16（2021）1–26。https://doi.org/10.1371/journal.pone.0247582。[18] Y. Akgül，A.O. Uymaz，高等教育中的 Facebook/Meta 使用：基于深度学习的双阶段结构方程模型-人工神经网络分析，美国施普林格，2022。https://doi.org/10.1007/s10639-022-11012-9。[19] P.S. Lo，Y.K. Dwivedi，G. Wei-Han Tan，K.B. Ooi，E. Cheng-Xi Aw，B. Metri，消费者为何在直播期间冲动购买？基于深度学习的双阶段结构方程模型-人工神经网络分析，《商业研究杂志》147（2022）325–337。https://doi.org/10.1016/j.jbusres.2022.04.013。[20] A. Fisher，C. Rudin，F. Dominici，所有模型都是错误的，但许多模型很有用：通过同时研究整个预测模型类别来学习变量的重要性，《机器学习研究杂志》20（2019）1–81。[21] L. Breiman，随机森林，《机器学习》45（2001）5–32。https://doi.org/10.1023/A:1010933404324。[22] J. Friedman，贪心函数逼近：梯度提升机，《统计学年鉴》29（2001）1189–1232。https://www.jstor.org/stable/2699986。[23] D.W. Apley，J. Zhu，黑箱监督学习模型中预测变量效应的可视化，《皇家统计学会杂志B辑：统计方法论》82（2020）1059–1086。https://doi.org/10.1111/rssb.12377。[24] P. Biecek，T. Burzykowski，《解释性模型分析》，CRC出版社，Boca Raton，佛罗里达州，2021。https://pbiecek.github.io/ema/。[25] D. Liu，S. Li，Y. Yu，I. Moustaki，有序变量间部分关联的评估：量化、可视化与假设检验，《美国统计学会杂志》116（2021）955–968。https://doi.org/10.1080/01621459.2020.1796394。[26] H.B. Mann，针对趋势的非参数检验，《计量经济学》13（1945）245–259。https://doi.org/10.2307/1907187。[27] M.G. Kendall，《秩相关方法》，Griffin，1948。[28] H. Baniecki，W. Kretowicz，P. Piatyszek，J. Wisniewski，P. Biecek，dalex：借助 Python 中的交互式可解释性与公平性实现负责任的机器学习，《机器学习研究杂志》22（2021）1–7。http://jmlr.org/papers/v22/20-1473.html。[29] L. Breiman，统计建模：两种文化，《统计科学》16（2001）199–215。https://doi.org/10.1214/ss/1009213726。[30] J.M. Collins，M.R. Clark，神经计算理论在职场行为预测中的应用：网络分析的示例与评估，《人事心理学》46（1993）503–524。https://doi.org/10.1111/j.1744-6570.1993.tb00882.x。[31] W.S. Sarle，人工神经网络与统计模型，载于：第九届年度 SAS 用户组国际会议论文集，1994：第

1538–1550. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.27.699.[32] Y. LeCun、Y. Bengio、G. Hinton，深度学习，《自然》521（2015）436–444。https://doi.org/10.1038/nature14539。[33] I. Goodfellow、Y. Bengio、A. Courville，《深度学习》，麻省理工学院出版社，马萨诸塞州剑桥，2016年。[34] X. Glorot、Y. Bengio，理解深度前馈神经网络训练的难点，载于：第十三届国际人工智能与统计学会议论文集，机器学习研究论文集，2010年：第249–256页。https://proceedings.mlr.press/v9/glorot10a.html。[35] G.E. Hinton、S. Osindero、Y.W. Teh，深度信念网络的快速学习算法，《神经计算》18（2006）1527–1554。https://doi.org/10.1162/neco.2006.18.7.1527。[36] B. Efron、T. Hastie，《计算机时代的统计推断：算法、证据与数据科学》，剑桥大学出版社，剑桥，2016年。https://doi.org/10.1017/cbo9781316576533。[37] G. Orrù、M. Monaro、C. Conversano、A. Gemignani、G. Sartori，机器学习在心理测量学与心理学研究中的应用，《心理学前沿》10（2020）1–10。https://doi.org/10.3389/fpsyg.2019.02970。[38] S. Papadokonstantakis、A. Lygeros、S.P. Jacobsson，神经网络中变量影响推断近期方法的比较，《神经网络》19（2006）500–513。https://doi.org/10.1016/j.neunet.2005.09.002。[39] P. Howes、N. Crook，利用输入参数影响来辅助前馈神经网络决策，《神经计算》24（1999）191–206。https://doi.org/10.1016/S0925-2312(98)00102-7。[40] L.I. Nord、S.P. Jacobsson，一种考察变量对计算神经网络模型贡献的新方法，《化学计量学与智能实验室系统》44（1998）153–160。https://doi.org/10.1016/S0169-7439(98)00118-X。[41] C. Molnar，《可解释机器学习》，第2版，慕尼黑，2022年。https://christophm.github.io/interpretable-ml-book/。[42] A.S. Selya、J.S. Rose、L.C. Dierker、D. Hedeker、R.J. Mermelstein，使用 PROC MIXED 计算科恩 f²（局部效应量度量）的实用指南，《心理学前沿》3（2012）1–6。https://doi.org/10.3389/fpsyg.2012.00111。[43] R. Bakeman、D. Mcarthur，如何确定有重复测量和无重复测量的多元回归分析的统计功效，《行为研究方法、仪器与计算机》31（1999）150–154。https://doi.org/10.3758/BF03207705。[44] B. Boehmke，使用 DALEX 实现模型可解释性，《加州大学商业分析 R 编程指南》（2018）。http://uc-r.github.io/dalex。[45] R.R. Wilcox，《现代统计方法基础》，第2版，施普林格，纽约州纽约市，2010年。[46] F. Wang、W. Shao、H. Yu、G. Kan、X. He、D. Zhang、M. Ren、G. Wang，重新评估 Mann-Kendall 检验检测水文气象时间序列单调趋势的效力，《地球科学前沿》8（2020）1–12。https://doi.org/10.3389/feart.2020.00014。[47] D.W. Meals、J. Spooner、S.A. Dressing、J.B. Harcum，单调趋势的统计分析，《国家非点源监测计划》（2011）1–23。https://www.epa.gov/sites/default/files/2016-05/documents/tech_notes_6_dec2013_trend.pdf。[48] K.H. Hamed、A.R. Rao，针对自相关数据的修正 Mann-Kendall 趋势检验，《水文杂志》204（1998）182–196。https://doi.org/10.1016/S0022-1694(97)00125-X。[49] M.M. Hussain、I. Mahmud，pyMannKendall：用于 Mann-Kendall 非参数趋势检验系列的 Python 软件包，《开源软件杂志》4（2019）1–3。https://doi.org/10.21105/joss.01556。[50] R.M. Hirsch、J.R. Slack、R.A. Smith，月度水质数据的趋势分析技术，《水资源研究》18（1982）107–121。https://doi.org/10.1029/WR018i001p00107。[51] D.G. Altman、J.M. Bland，统计学札记：如何根据置信区间得出 P 值，《英国医学杂志》343（2011）1–2。https://doi.org/10.1136/bmj.d2304。[52] Y.S. Güçlü，通过与经典 Mann-Kendall 检验和 ITA 对比改进趋势分析的可视化，《水文杂志》584（2020）124674。https://doi.org/10.1016/j.jhydrol.2020.124674。[53] A.N. Pettitt，一种针对变点问题的非参数方法，《英国皇家统计学会会刊 C辑（应用统计）》28（1979）126–135。https://doi.org/10.2307/2346729。[54] K. He、X. Zhang、S. Ren、J. Sun，深入探究整流器：在 ImageNet 分类上实现超越人类水平的性能，载于：IEEE 国际计算机视觉会议论文集，2014年：第1026–1034页。https://openaccess.thecvf.com/content_iccv_2015/html/He_Delving_Deep_into_ICCV_2015_paper.html。[55] M. Ek man，《学习深度学习：使用 TensorFlow 学习神经网络、计算机视觉、自然语言处理和 Transformer 的理论与实践》，Addison-Wesley，马萨诸塞州波士顿，2022年。[56] D.P. Kingma、J.L. Ba，ADAM：一种随机优化方法，载于：第三届国际学习表征会议（ICLR 2015）会议论文集，会议论文集， 加利福尼亚州圣迭戈，2015年：第1–15页。https://doi.org/10.48550/arXiv.1412.6980。[57] G.E. Hinton、N. Srivastava、A. Krizhevsky、I. Sutskever、R.R. Salakhutdinov，通过防止特征检测器的协同适应来改进神经网络，（2012）1–18。http://arxiv.org/abs/1207.0580。[58] J. Booz、J. McGiff、W.G. Hatcher、W. Yu、J. Nguyen、C. Lu，调整深度学习性能以用于 Android 恶意软件检测，载于：2018年 IEEE/ACIS 第19届软件工程、人工智能、网络与并行/分布式计算国际会议论文集（SNPD 2018），（2018）140–145。https://doi.org/10.1109/SNPD.2018.8441128。[59] ESS 1-9，《欧洲社会调查：累积文件》，ESS 1-9，数据文件版本1.0，NSD——挪威研究数据中心，挪威——数据档案分发机构，ESS 数据 ESS ERIC。（2020）。https://doi.org/10.21338/NSD-ESS-CUMULATIVE。[60] B. Mesquita、J. De Leersnyder、M. Boiger，情绪的文化心理学，载于：L. Feldman Barrett、M. Lewis、J.M. Haviland-Jones（编），《情绪手册》，第4版，吉尔福德出版社，纽约州纽约市，2016年：第393–411页。[61] R.A. Shweder、J. Haidt、R. Horton、C. Joseph，情绪的文化心理学：古老而焕新的领域，载于：M. Lewis、J.M. Haviland-Jones、L.F. Barrett（编），《情绪手册》，第3版，吉尔福德出版社，纽约州纽约市，2008年：第409–427页。[62] K.A. Lindquist、L.F. Barrett，人脑的功能架构：情绪科学的新兴见解，《认知科学趋势》16（2012）533–540。https://doi.org/10.1016/j.tics.2012.09.005。

[63] E. Diener、D. Kahneman、W. Tov、R. Arora，收入与生活判断而非感受之间的关联，载于：E. Diener、J.F. Helliwell、D. Kahneman（编），《国际幸福感差异》，牛津大学出版社，纽约州纽约市，2010年。[64] E. Diener，主观幸福感：幸福科学与建立国家指数的提议，《美国心理学家》55（2000）34–43。https://doi.org/10.1037/0003-066X.55.1.34。[65] E. Diener、S. Oishi、R.E. Lucas，人格、文化与主观幸福感：对生活的情感与认知评价，《心理学年度评论》54（2003）403–425。https://doi.org/10.1146/annurev.psych.54.101601.145056。[66] L.B. Aknin、E.W. Dunn、J.F. Helliwell、R. Biswas-Diener、P. Nyende、C.P. Barrington-Leigh、J. Burns、I. Kemeza、C.E. Ashton-James、M.I. Norton，亲社会支出与幸福感：心理普遍性的跨文化证据，《人格与社会心理学杂志》104（2013）635–652。https://doi.org/10.1037/a0031578。[67] P.H. van der Meer、R. Wielers，是什么让劳动者感到幸福？《应用经济学》45（2013）357–368。https://doi.org/10.1080/00036846.2011.602011。[68] P.A. Linley、S. Joseph，应用积极心理学：专业实践的新视角，载于：P.A. Linley、S. Joseph（编），《积极心理学实践》，约翰·威立父子出版公司，新泽西州霍博肯，2004年：第3–14页。[69] J.F. Helliwell、C. Barrington-Leigh、A. Harris、H. Huang，关于幸福感社会环境的国际证据，载于：E. Diener、J.F. Helliwell、D. Kahneman（编），《国际幸福感差异》，牛津大学出版社，纽约州纽约市，2010年：第291–27页。[70] W. Messner，人类情绪脑系统人工表征中的文化差异：一项深度学习研究，《国际营销学杂志》30（2022）21–43。https://doi.org/10.1177/1069031x221123993。[71] R. Veenhoven，幸福相关因素，世界幸福数据库，鹿特丹伊拉斯姆斯大学。（2019）。https://worlddatabaseofhappiness.eur.nl/hap_cor/cor_fp.htm（访问于2020年6月9日）。[72] R. Veenhoven，幸福，载于：A.C. Michalos（编），《生活质量与幸福感研究百科全书》，施普林格，多德雷赫特，2014年：第2637–2641页。https://doi.org/10.1007/978-94-007-0753-5。[73] E. Diener、E.M. Suh、R.E. Lucas、H.L. Smith，主观幸福感：三十年的进展，《心理学公报》125（1999）276–302。https://doi.org/10.1037/0033-2909.125.2.276。[74] W. Messner，主观测量对社会进步量化的贡献：来自欧洲和以色列的证据，《国际社会学与社会政策杂志》36（2016）258–268。https://doi.org/10.1108/IJSSP-06-2015-0060。[75] J.F. Helliwell、S. Wang，信任与幸福感，NBER工作论文系列 W15911。（2010）。https://www.nber.org/papers/w15911.pdf（访问于2020年6月1日）。[76] E.K. Kelloway、N. Turner、J. Barling、C. Loughlin，变革型领导与员工心理幸福感：员工对领导的信任的中介作用，《工作与压力》26（2012）39–55。https://doi.org/10.1080/02678373.2012.660774。[77] M. Kuhn、K. Johnson，《应用预测建模》，施普林格，纽约州纽约市，2016年。[78] R.D. Reed、R.J. Marks II，《神经网络锻造：前馈人工神经网络中的监督学习》，麻省理工学院出版社，马萨诸塞州剑桥，1999年。[79] Z.C. Lipton，模型可解释性的神话：在机器学习中，可解释性这一概念既重要又难以捉摸，《Queue》16（2018）1–28。https://doi.org/10.1145/3236386.3241340。[80] L.-V. Herm、K. Heinrich、J. Wanner、C. Janiesch，别再依据可解释性对机器学习算法排序！一项以用户为中心的性能与可解释性研究，《国际信息管理杂志》（2022）。https://doi.org/10.1016/j.ijinfomgt.2022.102538。[81] D. Chicco、M.J. Warrens、G. Jurman，在回归分析评估中，决定系数 R-squared 比 SMAPE、平均绝对误差、平均绝对百分比误差、均方误差和均方根误差更具信息量，《PeerJ计算机科学》7（2021）1–24。https://doi.org/10.7717/PEERJ-CS.623。[82] J. Ish-Horowicz、D. Udwin、S. Flaxman、S. Filippi、L. Crawford，通过变量重要性解读深度神经网络，《机器学习研究杂志》21（2019）1–30。[83] R.D. De Veaux、L.H. Ungar，多重共线性：两个非参数回归的故事，载于：P. Cheeseman、R.W. Oldford（编），《从数据中选择模型》，《统计学讲义》，施普林格，纽约州纽约市，1994年：第393–403页。https://doi.org/10.1007/978-1-4612-2660-4_40。[84] J.Y. Chan、S. Mun、H. Leow、K.T. Bea、W.K. Cheng、S.W. Phoong、Z. Hong、J. Lin、Y. Chen，一种用于缓解多重共线性的相关性嵌入注意力模块：算法交易应用，《数学》10（2022）1–13。https://doi.org/10.3390/math10081231。[85] M. Paliwal、U.A. Kumar，评估变量对前馈神经网络的贡献，《应用软计算》11（2011）3690–3696。https://doi.org/10.1016/j.asoc.2011.01.040。