面向深度神经网络解释的第一性原理：基于等效交互理论解析学习动态性

周慧琳[1]，任启涵[1]，张俊鹏[1]，张拳石[1,2]
[1]上海交通大学电子信息与电气工程学院，中国上海市，200240
[2]上海交通大学计算机学院，中国上海市，200240

摘要：当前关于深度学习可解释性的大部分研究都是经验主义的，而是否存在第一性原理，从不同角度全方位严谨解释深度神经网络的内在机理，成为可解释人工智能领域亟待解决的核心科学问题之一。本文探讨等效交互理论可否用于深度神经网络的第一性原理解释分析。我们认为，该理论之所以具备较强的解释能力，主要体现在以下 4 个方面：（1）建立了一套新的公理体系，将深度神经网络的决策逻辑转化为一系列符号化的交互；（2）能够同时解释深度学习的多种典型特征，包括网络的泛化能力、抗敏感性、表征瓶颈以及学习动态性；（3）提供了统一解释深度学习算法的数学工具，从而能够系统地解释各种经验归因方法以及对抗迁移性方法背后的机制；（4）分析深度神经网络建模过程中交互复杂度的双阶段动态变化，解释深度神经网络在训练过程中建模的复杂性以及泛化能力和抗敏感性之间的联系，从而深入揭示深度神经网络的泛化能力和抗敏感性在学习阶段的内在变化机理。

关键词：第一性原理解释；等效交互理论；双阶段动态交互；学习动态性
https://doi.org/10.1631/FITEE.2401025