

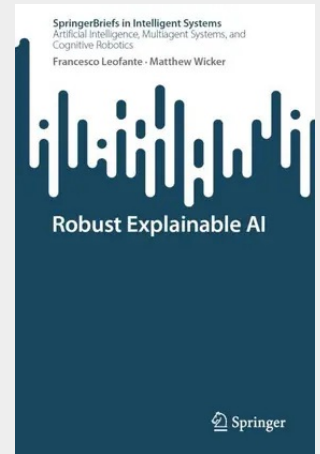
Robust Explainable AI

The area of Explainable Artificial Intelligence (XAI) is concerned with providing methods and tools to improve the interpretability of black-box learning models. While several approaches exist to generate explanations, they are often lacking robustness, e.g., they may produce completely different explanations for similar events. This phenomenon has troubling implications, as lack of robustness indicates that explanations are not capturing the underlying decision-making process of a model and thus cannot be trusted.

This book aims at introducing Robust Explainable AI, a rapidly growing field whose focus is to ensure that explanations for machine learning models adhere to the highest robustness standards. We will introduce the most important concepts, methodologies, and results in the field, with a particular focus on techniques developed for feature attribution methods and counterfactual explanations for deep neural networks.

As prerequisites, a certain familiarity with neural networks and approaches within XAI is desirable but not mandatory. The book is designed to be self-contained, and relevant concepts will be introduced when needed, together with examples to ensure a successful learning experience.

The area of Explainable Artificial Intelligence (XAI) is concerned with providing methods and tools to improve the interpretability of black-box learning models. While several approaches exist to generate explanations, they are often lacking robustness, e.g., they may produce completely different explanations for similar events. This phenomenon has troubling implications, as lack of robustness indicates that explanations are not capturing the underlying decision-making process of a model and thus cannot be trusted. This book aims at introducing Robust Explainable AI, a rapidly growing field whose focus is to ensure that explanations for machine learning models adhere to the highest robustness standards. We will introduce the most important concepts, methodologies, and results in the field, with a particular focus on techniques developed for feature attribution methods and counterfactual explanations for deep neural networks. As prerequisites, a certain familiarity with neural networks and approaches within XAI is desirable but not mandatory. The book is designed to be self-contained, and relevant concepts will be introduced when needed, together with examples to ensure a successful learning experience.



53,49 €
49,99 € (zzgl. MwSt.)

Lieferfrist: bis zu 10 Tage

Artikelnummer: 9783031890215
Medium: Buch
ISBN: 978-3-031-89021-5
Verlag: Palgrave Macmillan
Erscheinungstermin: 25.05.2025
Sprache(n): Englisch
Auflage: Erscheinungsjahr 2025
Serie: SpringerBriefs in Intelligent Systems
Produktform: Kartoniert
Gewicht: 143 g
Seiten: 71
Format (B x H): 155 x 235 mm

