1. Record Nr. UNISA996547963703316

Autore	Huang Xiaowei
Titolo	Machine Learning Safety / / Xiaowei Huang, Gaojie Jin, and Wenjie Ruan
Pubbl/distr/stampa	Singapore : , : Springer Nature Singapore Pte Ltd., , [2023] ©2023
ISBN	9789811968143 9789811968136
Edizione	[1st ed. 2023.]
Descrizione fisica	1 online resource (319 pages)
Collana	Artificial Intelligence: Foundations, Theory, and Algorithms Series
Disciplina	005.8
Soggetti	Computer security Machine learning - Safety measures
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di bibliografia	Includes bibliographical references.
Nota di contenuto	1. Introduction -- 2. Safety of Simple Machine Learning Models -- 3. Safety of Deep Learning -- 4. Robustness Verification of Deep Learning -- 5. Enhancement to Robustness and Generalization -- 6. Probabilistic Graph Model -- A. Mathematical Foundations -- B. Competitions.
Sommario/riassunto	Machine learning algorithms allow computers to learn without being explicitly programmed. Their application is now spreading to highly sophisticated tasks across multiple domains, such as medical diagnostics or fully autonomous vehicles. While this development holds great potential, it also raises new safety concerns, as machine learning has many specificities that make its behaviour prediction and assessment very different from that for explicitly programmed software systems. This book addresses the main safety concerns with regard to machine learning, including its susceptibility to environmental noise and adversarial attacks. Such vulnerabilities have become a major roadblock to the deployment of machine learning in safety-critical applications. The book presents up-to-date techniques for adversarial attacks, which are used to assess the vulnerabilities of machine learning models; formal verification, which is used to determine if a trained machine learning model is free of vulnerabilities; and adversarial training, which is used to enhance the training process and reduce vulnerabilities. The book aims to improve readers' awareness of

the potential safety issues regarding machine learning models. In addition, it includes up-to-date techniques for dealing with these issues, equipping readers with not only technical knowledge but also hands-on practical skills.