1. Record Nr.	UNISA996466320103316
Titolo	Explainable AI: Interpreting, Explaining and Visualizing Deep Learning [[electronic resource] /] / edited by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, Klaus-Robert Müller
Pubbl/distr/stampa	Cham : , : Springer International Publishing : , : Imprint : Springer, , 2019
ISBN	3-030-28954-0
Edizione	[1st ed. 2019.]
Descrizione fisica	1 online resource (XI, 439 p. 152 illus., 119 illus. in color.)
Collana	Lecture Notes in Artificial Intelligence ; ; 11700
Disciplina	006.32
Soggetti	Artificial intelligence Optical data processing Computers Computer security Computer organization Artificial Intelligence Image Processing and Computer Vision Computing Milieux Systems and Data Security Computer Systems Organization and Communication Networks
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di bibliografia	Includes bibliographical references and index.
Nota di contenuto	Towards Explainable Artificial Intelligence -- Transparency: Motivations and Challenges -- Interpretability in Intelligent Systems: A New Concept? -- Understanding Neural Networks via Feature Visualization: A Survey -- Interpretable Text-to-Image Synthesis with Hierarchical Semantic Layout Generation -- Unsupervised Discrete Representation Learning -- Towards Reverse-Engineering Black-Box Neural Networks -- Explanations for Attributing Deep Neural Network Predictions -- Gradient-Based Attribution Methods -- Layer-Wise Relevance Propagation: An Overview -- Explaining and Interpreting LSTMs -- Comparing the Interpretability of Deep Networks via Network Dissection -- Gradient-Based vs. Propagation-Based Explanations: An Axiomatic Comparison -- The (Un)reliability of Saliency Methods --

Sommario/riassunto

The development of “intelligent” systems that can take decisions and perform autonomously might lead to faster and more consistent decisions. A limiting factor for a broader adoption of AI technology is the inherent risks that come with giving up human control and oversight to “intelligent” machines. Forsensitive tasks involving critical infrastructures and affecting human well-being or health, it is crucial to limit the possibility of improper, non-robust and unsafe decisions and actions. Before deploying an AI system, we see a strong need to validate its behavior, and thus establish guarantees that it will continue to perform as expected when deployed in a real-world environment. In pursuit of that objective, ways for humans to verify the agreement between the AI decision structure and their own ground-truth knowledge have been explored. Explainable AI (XAI) has developed as a subfield of AI, focused on exposing complex AI models to humans in a systematic and interpretable manner. The 22 chapters included in this book provide a timely snapshot of algorithms, theory, and applications of interpretable and explainable AI and AI techniques that have been proposed recently reflecting the current discourse in this field and providing directions of future development. The book is organized in six parts: towards AI transparency; methods for interpreting AI systems; explaining the decisions of AI systems; evaluating interpretability and explanations; applications of explainable AI; and software for explainable AI.