ellis unit | TURIN Talk

Politecnico di Torino

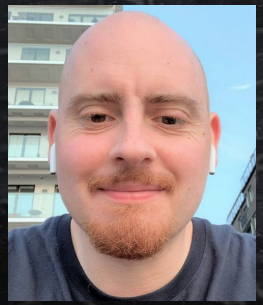
ARTIFICIAL INTELLIGENCE HUB
POLITECNICO DI TORINO

VANDAL
http://vandal.polito.it

www.ellis.eu

Francesco Quinzan

Computer Science Department
University of Oxford



Jan. 14th, 2025
starting at 14:30 PM
Online - link will be
distributed soon!

AI Safety: Challenges and Opportunities

Recent successes of AI and Machine Learning have ignited a fast transfer of technology from research into products and government services. This phenomenon has created a range of problems, which can be broadly attributed to the interaction between technology and society. Examples of these problems are bias and unfairness, lack of robustness, and lack of transparency. In this talk, I will discuss some of the main challenges in AI Safety, focusing on safety-critical applications. I will argue that it is possible to design AI systems that are robust and capable of generalizing effectively, by uncovering the causal mechanisms of the underlying data generating process. I will illustrate recent advancements in this field and discuss possible future directions.

Francesco Quinzan is an associate researcher at the Computer Science department at the University of Oxford, hosted by Marta Kwiatkowska. Previously, he was a postdoc at the Division of Decision and Control Systems at KTH Royal Institute of Technology, where he worked with Stefan Bauer and Cristian Rojas. Francesco obtained his Ph.D. in Computer Science in 2022 from the Hasso Plattner Institute in Germany, where he was advised by Tobias Friedrich. During his studies, he visited various institutes, including the Nanjing University, and the at the Max Plank Institute for Intelligent Systems, where he was hosted by the group of Bernhard Schölkopf. Francesco studied mathematics at the University of Roma Tre, where he graduated with honors.