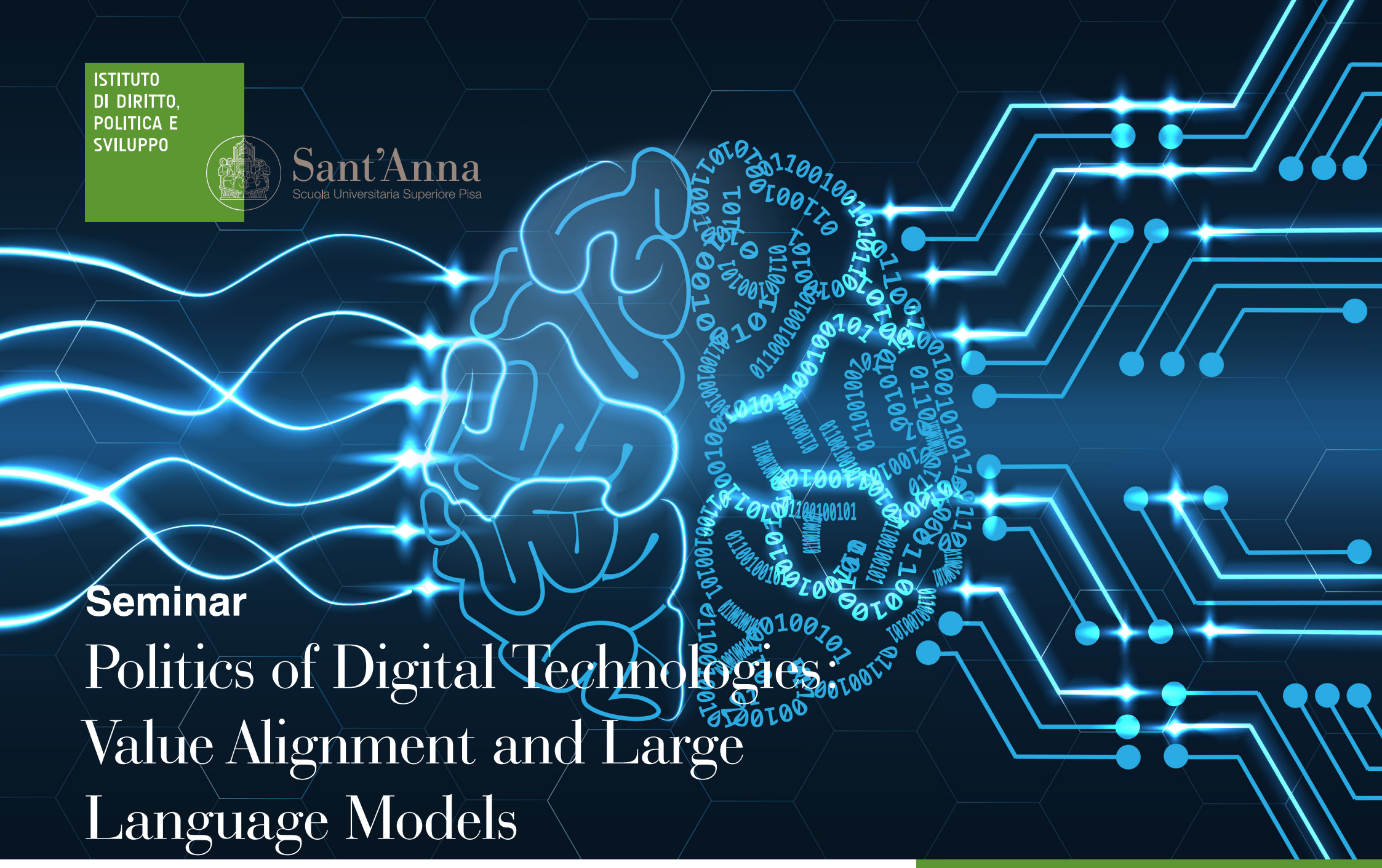
ISTITUTO
DI DIRITTO,
POLITICA E
SVILUPPO

Sant'Anna
Scuola Universitaria Superiore Pisa



Seminar

Politics of Digital Technologies: Value Alignment and Large Language Models

Abstract

The value alignment problem in Artificial Intelligence (AI) asks the question "How should we align AI systems with human values?" In this talk, we'll explore the value alignment problem with respect to large language models, such as ChatGPT. The keynote speaker suggests two conditions for successful alignment and argue that current large language models do not satisfy one of them. Finally, she will draw some important practical implications of her proposal and sketch future directions of research.

Introduction and chair

Barbara Henry, full professor of political philosophy at Dripolis Institute, Sant'Anna School

Speaker

Prof. **Aloosa Kasirzadeh** is a philosopher, mathematician, and systems engineer. She is an assistant professor (Chancellor's Fellow) in the philosophy department and the Director of Research at the Centre for Technomoral Futures at the University of Edinburgh, and a Research Lead at the Alan Turing Institute. Prior to this, she held research positions at DeepMind and Australian National University. She has a Ph.D. in philosophy of science and technology (2021) from the University of Toronto and a Ph.D. in mathematics (2015) from the Ecole Polytechnique of Montreal. Her current research is focused on ethics, safety, and philosophy of AI (value alignment, interpretability, generative models, recommender systems) and philosophy of science (explanation, prediction, complex systems, automating science). She also collaborates with public and private institutions as an advisor.

Date
16-17/10/2023
Hour
10-12:30
Where
Aula 5, Palazzo Maffi - Pisa
Streaming live on WebEx Platform
Link for Part I (16 October):
Click here

Link for Part II (17 October)
Click here

Scan the QR code to connect to the event web page