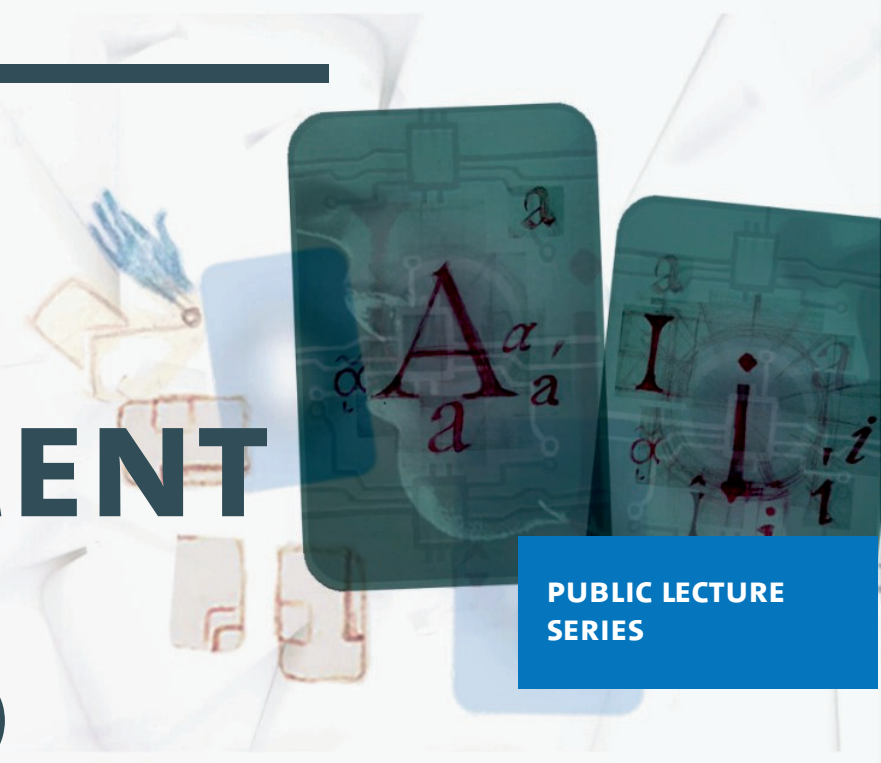
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

TAMING THE MACHINES

A FALLIBILIST APPROACH TO AI VALUE ALIGNMENT



PUBLIC LECTURE SERIES

PROF. DR. IBO VAN DE POEL (DELFT UNIVERSITY OF TECHNOLOGY, NL)

About the lecture

Value alignment is important to ensure that AI systems remain aligned with human intentions, preferences, and values. It has been suggested that it can best be achieved by building AI systems that can track preferences or values in real-time. In my talk, I argue against this idea of real-time value alignment. First, I show that the value alignment problem is not unique to AI, but applies to any technology, thus opening up alternative strategies for attaining value alignment. Next, I argue that due to uncertainty about appropriate alignment goals, real-time value alignment may lead to harmful optimization and therefore will likely do more harm than good. Instead, it is better to base value alignment on a fallibilist epistemology, which assumes that complete certainty about the proper target of value alignment is and will remain impossible. Three alternative principles for AI value alignment are proposed: 1) adopt a fallibilist epistemology regarding the target of value alignment; 2) focus on preventing serious misalignments rather than aiming for perfect alignment; 3) retain AI systems under human control even if it comes at the cost of full value alignment.

Monday, 27. January 2025
18:15-19:45 (CET)

Flügelbau Ost, 2. OG, Raum O 221
Edmund-Siemers-Allee 1
20146 Hamburg

If you like to join us virtually, register at uhh.de/inf-eit.

ETHIK IN DER
INFORMATIONSTECHNOLOGIE

Kontakt: ttm.inf@uni-hamburg.de