Interdisziplinärer Workshop

AI Alignment

Perspectives in Law and Ethics

12. und 13.11.2026

Universität Tübingen, Center for Interdisciplinary and Intercultural Studies Villa Köstlin,

Rümelinstr. 27, 72070 Tübingen

Organisatoren: Dr. Hans Lind, PD Dr. Jörg Noller

gefördert durch die

Fritz Thyssen Stiftung
für Wissenschaftsförderung

Künstliche Intelligenz (KI) durchdringt zunehmend gesellschaftlich sensible Bereiche wie das Gesundheitswesen, das Finanzwesen und die öffentliche Verwaltung. Da KI-Systeme jedoch keine autonomen moralischen oder rechtlichen Subjekte sind, bedarf es des sogenannten KI-Alignments: des komplexen Prozesses, maschinelle Systeme so zu konfigurieren, dass sie menschlichen Werten und rechtlichen Normen entsprechen. Dieser Prozess erweist sich aus ethischer wie rechtlicher Sicht als überaus komplex. Zum einen steht die Frage nach denjenigen Normen im Zentrum, an denen KI ausgerichtet werden soll. Zum anderen steht der ethische und rechtliche Prozess der Ausrichtung der KI-Systeme selbst im Zentrum, wobei hier ethische und technische Fragen aufs Engste zusammenhängen.

Der Workshop positioniert sich explizit im internationalen Forschungsfeld zu *AI Alignment*, das bislang maßgeblich von technisch orientierten Debatten um AI Safety, Machine Learning Alignment und Governance-Modelle geprägt ist. Während bislang vor allem Fragen der Risikominimierung, Robustheit und Regulierung dominieren, setzt der Workshop einen eigenständigen Akzent, indem er Alignment systematisch als normativen, epistemischen und institutionellen Prozess analysiert. Damit schließt er an internationale Diskurse zu *trustworthy AI*, zur Regulierung generativer Modelle sowie zur Theorie algorithmischer Verantwortung an, erweitert diese jedoch um eine grundlagentheoretische Perspektive auf Moralität, Rechtssubjektivität und praktische Vernunft.

Der Workshop untersucht insofern KI-Alignment nicht als rein technisches Problem, sondern als relationalen und prozeduralen Prozess im Sinne der Mensch-Maschine-Interaktion. Ziel ist es, Alignment als dynamische Form der Mensch-Maschine-Interaktion zu verstehen, die tief in sozialen, historischen und institutionellen Kontexten verwurzelt ist. Das Vorhaben bringt Expertinnen und Experten aus der Philosophie und den Rechtswissenschaften zusammen, um die normativen, epistemischen und technischen Grundlagen der KI-Alignment-Praktiken kritisch und interdisziplinär zu reflektieren.