The Pinocchio architecture for human-AI interaction

Leon van der Torre (joint work with Beishui Liao and Marija Slavkovik)

An autonomous system is constructed by a manufacturer, operates in a society subject to norms and laws, and is interacting with end-users. We address the challenge of how the moral values and views of all stakeholders can be integrated and reflected in the moral behaviour of the autonomous system. We propose an artificial moral agent architecture that uses techniques from normative systems and formal argumentation to reach moral agreements among stakeholders. We show how our architecture can be used not only for ethical practical reasoning and collaborative decision-making,but also for the explanation of such moral behavior.

Reference

Beishui Liao, Marija Slavkovik and Leendert van der Torre. Building Jiminy Cricket:
An Architecture for Moral Agreements Among Stakeholders. Proceedings of the Second AAAI / ACM conference on artificial intelligence, ethics and society (AIES 2019), 2019.