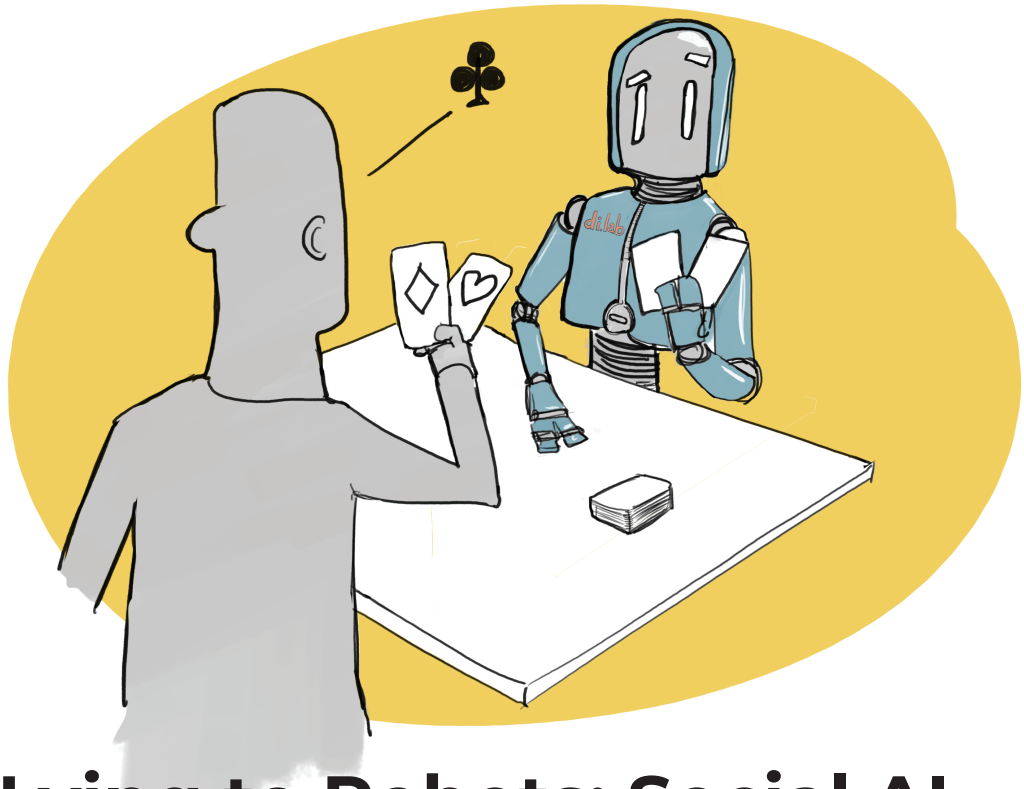
di.lab



Lying to Robots: Social AI Deception Awareness & Deterrence

What should robots do when they are being lied to? Join us in investigating computational models for handling human deception. As AI agents become more prevalent, it is paramount that we design for a human propensity for exploiting and corrupting these systems.

In this project, you will design, develop, and test mechanisms (protocols and modalities) for AI agents within AI-human interaction where the human party is incentivised to be dishonest. You will work with PhDs within Designing Intelligence Lab to develop a framework for detecting and a library of mechanisms deterring deception through conversational interfaces.

#HumanComputerInteraction #BehaviourChange

Catherine Oertel: C.R.M.M.Oertel@tudelft.nl
Eric (Heng) Gu: h.gu@tudelft.nl

Sound like a project that suits your interests and skills? Let's set up a meeting to get to know each other. Learn more about us at: www.di-lab.space