

Workshop: Opportunities and Limits of Artificial Moral Agents

Day 1 (5 March)	Location: Parnassos Cultuurcentrum, Kruisstraat 201
12-12:15	Opening
12:15-13:00	Henry Prakken (Utrecht University): (AI &) Law and (Machine) Ethics
<p><i>Abstract:</i> In this talk I will compare law and ethics, with particular focus on legal versus moral reasoning and decision-making. I will then give an overview of AI & Law research on legal reasoning and decision-making and how it could be relevant for machine ethics. Among other things I will discuss how AI & law formalisms can model and combine rule-following, balancing and particularist approaches to (machine) ethics.</p>	
13:00-14:00	Lunch
14:00-14:45	Eva Schmidt (TU Dortmund): The Reasons of AI Systems
<p><i>Abstract:</i> We argue that it can sometimes be appropriate to explain the outputs of artificial intelligent (AI) systems by appeal to the practical reasons of these systems, and that such explanations can be faithful—not mere convenient fictions, but genuine descriptions of what drives system outputs. (Joint work with Kevin Baum and Timo Speith.)</p>	
14:45-15:30	Emily Sullivan (University of Edinburgh): Can LLMs provide moral testimony?
<p><i>Abstract:</i> In this talk I raise a dilemma that either LLMs cannot provide moral testimony or if they do provide moral testimony, we still cannot gain moral understanding from LLMs. This dilemma makes us question what we should consider testimony and whether we can gain understanding through testimony alone.</p>	
15:30-16:00	Coffee break
16:00-16:45	Jan Broersen (Utrecht University): On the biases LLMs cannot have
<p>Cognitive scientists distinguish hundreds of different human biases, typically understood as systematic deviations from norms of rational judgment or decision-making. Much of the AI literature warns that artificial systems may replicate or even amplify such biases. In this talk, I argue that large language models in fact avoid a particular class of biases that are among the most harmful in human reasoning, and that this absence is connected to the distinctive way in which LLMs generate behaviour without engaging in deliberative agency. The result is a striking tension: the systems most often criticized for reproducing human bias may systematically lack some of the biases that most deeply distort human decision-making.</p>	
16:45-17:30	Drinks
Day 2 (6 March)	Location: Drift 27, 0.72
09:00-09:45	Aleks Knoks (University of Luxembourg): Metanormative Theory for RL-Based Moral Agents
<p><i>Abstract:</i> The talk has two related goals. The first is to draw out some ideas from recent metanormative theory that can directly inform research on machine ethics and value alignment. The second is to examine recent approaches to designing moral agents with reinforcement learning through the lens of these metanormative ideas.</p>	
09:45-10:30	Ibo van de Poel (TU Delft): AI, values and alignment

Abstract: How do we ensure that AI systems remain aligned with human values? Recently, worries have been expressed that the autonomy and increased ‘intelligence’ of AI systems may lead to systems that get out of control and are harmful to humans. This has led to proposals for AI systems that ensure value alignment by tracking human values or preferences in real-time (for example, Stuart Russell in his book *Human compatible: artificial intelligence and the problem of control*). In my talk I will criticize such proposals for being insufficient for addressing the alignment problem.

10:30-10:45	Closing
10:45-11:30	Coffee