Knowledge and Reasoning in Ethically Aligned AI Systems

Abstract: AI systems are increasingly acting in the same environment as humans, in areas as diverse as driving, assistive technology, and health care. Humans and machines will often need to work together and agree on common decisions. This requires alignment around some common moral values and ethical principles. Humans will accept and trust more machines that behave at least as ethically as other humans in the same environment. Moreover, shared moral values and ethical principles will help in collective human-machine decision making, where consensus or compromise is needed. In this scenario, end-to-end machine learning does not seem to provide all the necessary capabilities (such as explainability, value alignment, and fairness) to build trust in AI systems, and should be suitably combined with KR techniques that can help in modelling ethical priorities and reason with them. In this talk I will present some challenges in designing ethically aligned AI systems that humans can trust, and describe possible technical solutions for such challenges, based on a combination of machine learning techniques with knowledge modelling and reasoning components.