Invited Talk: Between Rules and Reasoning: Towards Machine Morality

Elizaveta Tennant[1,2]

[1]University College London, UK
[2]Google DeepMind, UK

Abstract

Is "learning" sufficient for machine morality, or are we teaching AI to mimic behavior without understanding the stakes? As AI enters increasingly sensitive social spheres, the tension between bottom-up learning and top-down formal constraints has become the central challenge in creating intrinsically aligned intelligent systems.

This talk maps the landscape of methodologies for developing AI morality along a continuum—from systems that infer ethics through data to those governed by rigid formal logic. By drawing parallels to the broader history of AI research, I will explore why neither extreme has succeeded in isolation. I will present a concrete framework for "hybrid morality," demonstrating how formalization can be paired with reinforcement learning to create LLM agents that don't just follow rules, but internalize moral goals. I will conclude with a look at the next frontier, reviewing recent proposals to combine Generative AI with verifiable and formal methods to ensure the systems of tomorrow are as safe as they are capable.

Invited Speaker's Bio

Elizaveta Tennant is an AI researcher specializing in AI alignment, machine morality, and intelligent agents. She is a PhD candidate in Computer Science at University College London (UCL), where she is affiliated with the Machine Intelligence Lab and studies how artificial agents can be guided by human values, moral reasoning, and social norms, particularly through reinforcement learning and large language model–based agents. Her academic background is interdisciplinary: she studied psychology and linguistics as an undergraduate at UCL and later conducted research in computational social science and political psychology at the University of Cambridge. While she has flipped her research focus into ML-first techniques in her PhD, she continues to take inspiration from the human sciences and humanities in formalising ethical alignment for AI. Her research has been presented at leading international venues, including ICLR and the AAAI/ACM Conference on AI, Ethics, and Society, and addresses topics such as social dilemmas, intrinsic moral rewards, and alignment in multi-agent settings. In 2025, she joined Google DeepMind as a Student Researcher, where she is continuing to focus on the responsible development of morally aligned AI systems.

Declaration on Generative AI

During the preparation of this manuscript, the author did not use any AI tools.

Machine Ethics: from formal methods to emergent machine ethics Workshop, January 26, AAAI 2026, Singapore
l.karmannaya.16@ucl.ac.uk (E. Tennant)
0000-0003-4833-6753 (E. Tennant)