Preface for The Machine Ethics Workshop: From Formal Methods to Emergent Machine Ethics

Hiroshi Yamakawa[1,2], Marija Slavkovik[3], Rafal Rzepka[4,*], Taichiro Endo[5,6], Louise A. Dennis[7], Raynaldio Limarga[7], Michael Fisher[7] and Ryutaro Ichise[8]

[1]*The University of Tokyo, Tokyo, Japan*
[2]*AI Alignment Network, Tokyo, Japan*
[3]*University of Bergen, Norway*
[4]*Faculty of Information Science and Technology, Hokkaido University, Sapporo, Japan*
[5]*Tokyo Gakugei University, Tokyo, Japan*
[6]*Kaname Project Co. Ltd., Aichi, Japan*
[7]*Department of Computer Science, University of Manchester, UK*
[8]*Institute of Science Tokyo, Tokyo, Japan*

Overview and Goals

Machine ethics is concerned with the behaviour of machines towards people and other machines. Advances in Artificial Intelligence (AI) continue to reinforce the need for research work in this field. In one direction this work involves specifying, implementing, and verifying ethical and safe AI. In another direction is the study of ethics that emerges by self-organizing inside AI-centric societies. We need to be able to verify the existing ethical capabilities of AI systems, and make advances towards understanding how transparent, collaboratively guidable AI ecosystems might evolve. All of these efforts, while happening at different technology and conceptual levels, are united in the need to foster international collaboration on AI-safety, multi-agent cooperation, and governance research.

Twenty years ago, a fall AAAI Fall symposium on Machine Ethics, kick-started the machine ethics field in computer science (https://auld.aaai.org/Library/Symposia/Fall/fs05-06.php). The time is right for a formative venue of a new generation of machine ethics researchers and teams. Building on emerging work in communities such as Japan's SIG-AGI (https://www.sig-agi.org/sig-agi/event/sig-agi-30-panel-en/), where foundations for Emergent Machine Ethics are being actively developed, this workshop marks a critical evolution from traditional top-down approaches to include bottom-up emergence, recognizing that as AI systems become increasingly autonomous, we need both formal verification methods and emergent approaches working in tandem to respect their agency while ensuring beneficial outcomes.

This workshop was intended to provide an arena for presenting new work on both classical topics studied in machine ethics and emergent machine ethics. To our knowledge, this is the first international workshop explicitly bridging formal methods and emergent approaches in machine ethics.

The Machine Ethics Workshop: From Formal Methods to Emergent Machine Ethics, January 26, AAAI 2026, Singapore
*Corresponding author.
hymkw@weblab.t.u-tokyo.ac.jp (H. Yamakawa); marija.slavkovik@uib.no (M. Slavkovik); rzepka@ist.hokudai.ac.jp (R. Rzepka); taichiro@u-gakugei.ac.jp (T. Endo); louise.dennis@manchester.ac.uk (L. A. Dennis); raynaldio.limarga@manchester.ac.uk (R. Limarga); michael.fisher@manchester.ac.uk (M. Fisher); ichise@iee.e.titech.ac.jp (R. Ichise)
https://www.aialign.net/team/ (H. Yamakawa); http://kabura.info/ (R. Rzepka); https://www.ai.iee.e.titech.ac.jp/ichise/ (R. Ichise)
0000-0002-6981-0349 (H. Yamakawa); 0000-0003-2548-8623 (M. Slavkovik); 0000-0002-8274-0875 (R. Rzepka); 0000-0003-1426-1896 (L. A. Dennis); 0009-0002-1142-7651 (R. Limarga); 0000-0002-0875-3862 (M. Fisher); 0000-0001-8474-0150 (R. Ichise)

Technical Program and Submissions

The workshop received 23 submissions from researchers worldwide. After a peer-review process, 11 papers were accepted for presentation (9 long and 2 short). One long paper was excluded from the proceedings as the authors did not attend the workshop.

The technical program reflects the diverse landscape of modern machine ethics research. Key themes include: formal specification and verification of ethical reasoning, computational modeling of morality, emergent ethics in multi-agent systems, value alignment through internal emergence, and human-AI co-creative guidance.

An introductory overview of the Emergent Machine Ethics (EME) framework was also presented to establish common ground among participants from diverse research backgrounds. The workshop concluded with a panel discussion (approximately 50 minutes), in which participants explored topics such as the unintended loss of pro-social flexibility through technology deployment, the democratization of agent value specification by end users, and the role of post-hoc rationalization in both human moral judgment and LLM chain-of-thought reasoning.

Invited Talk

We were honored to host two invited talks.

The first talk was by **Mizuki Oka** (Founder/Managing Director, Artificial Life Institute) titled *"From Unilateral Control to Social Homeostasis: Organic Alignment via Collective Predictive Coding"*. This talk introduced Organic Alignment, where humans and AI co-create shared meanings through Collective Predictive Coding and collective free energy minimization. Current AI alignment methods impose static ethics unilaterally, failing to accommodate diverse and evolving human values. The Organic Alignment approach shifts from hierarchical control to cultivating emergent ecosystems that support diverse world-views and mutual care.

The second talk was by **Elizaveta Tennant** (Google DeepMind (Student Researcher) & University College London (PhD)) titled *"Between Rules and Reasoning: Towards Machine Morality"*. This talk presented a hybrid morality framework combining formalization with reinforcement learning, enabling LLM agents to internalize moral goals rather than merely follow rules. Learning alone is insufficient for machine morality, as AI may mimic behavior without understanding moral stakes, creating tension between bottom-up learning and top-down formal constraints. The approach bridges data-driven ethics and formal logic, proposing to combine Generative AI with verifiable methods for safe and capable systems.

Acknowledgements

We would like to express our deepest gratitude to the Program Committee for their invaluable time and detailed feedback. We also thank the authors for their high-quality contributions and the participants for their engaging discussions that continue to drive this community forward. The workshop was made possible by the Distinguished International Associate Grant from the Royal Academy of Engineering of the UK. Further information, including presentation materials, is available at https://www.aialign.net/ws-machine-ethics/.

Organizers

Louise Dennis, University of Manchester, UK
Taichiro Endo, Tokyo Gakugei University, Japan / Kaname Project Co. Ltd., Japan
Michael Fisher, University of Manchester, UK
Ryutaro Ichise, Institute of Science Tokyo, Japan

Raynaldio Limarga, University of Manchester, UK
Rafal Rzepka, Hokkaido University, Japan
Marija Slavkovik, University of Bergen, Norway
Hiroshi Yamakawa, University of Tokyo, Japan / AI Alignment Network, Japan

Program Committee

- Kevin Baum, German Research Center for Artificial Intelligence (DFKI)
- Rémy Chaput, CPE Lyon, France
- Marina De Vos, University of Bath, UK
- Louise Dennis, University of Manchester, UK
- Taichiro Endo, Tokyo Gakugei University, Japan
- John-Stewart Gordon, Kaunas University of Technology, Lithuania
- Ryutaro Ichise, Institute of Science Tokyo, Japan
- Aleks Knoks, University of Luxembourg, Luxembourg
- Simon Kolker, University of Manchester, UK
- Raynaldio Limarga, University of Manchester, UK
- Nicholas Mattei, Tulane University, USA
- Vivek Nallur, University College Dublin, Ireland
- Rafal Rzepka, Hokkaido University, Japan
- Marija Slavkovik, University of Bergen, Norway
- Masashi Takeshita, Hokkaido University, Japan
- Hiroshi Yamakawa, University of Tokyo, Japan / AI Alignment Network, Japan
- Liuwen Yu, University of Luxembourg, Luxembourg