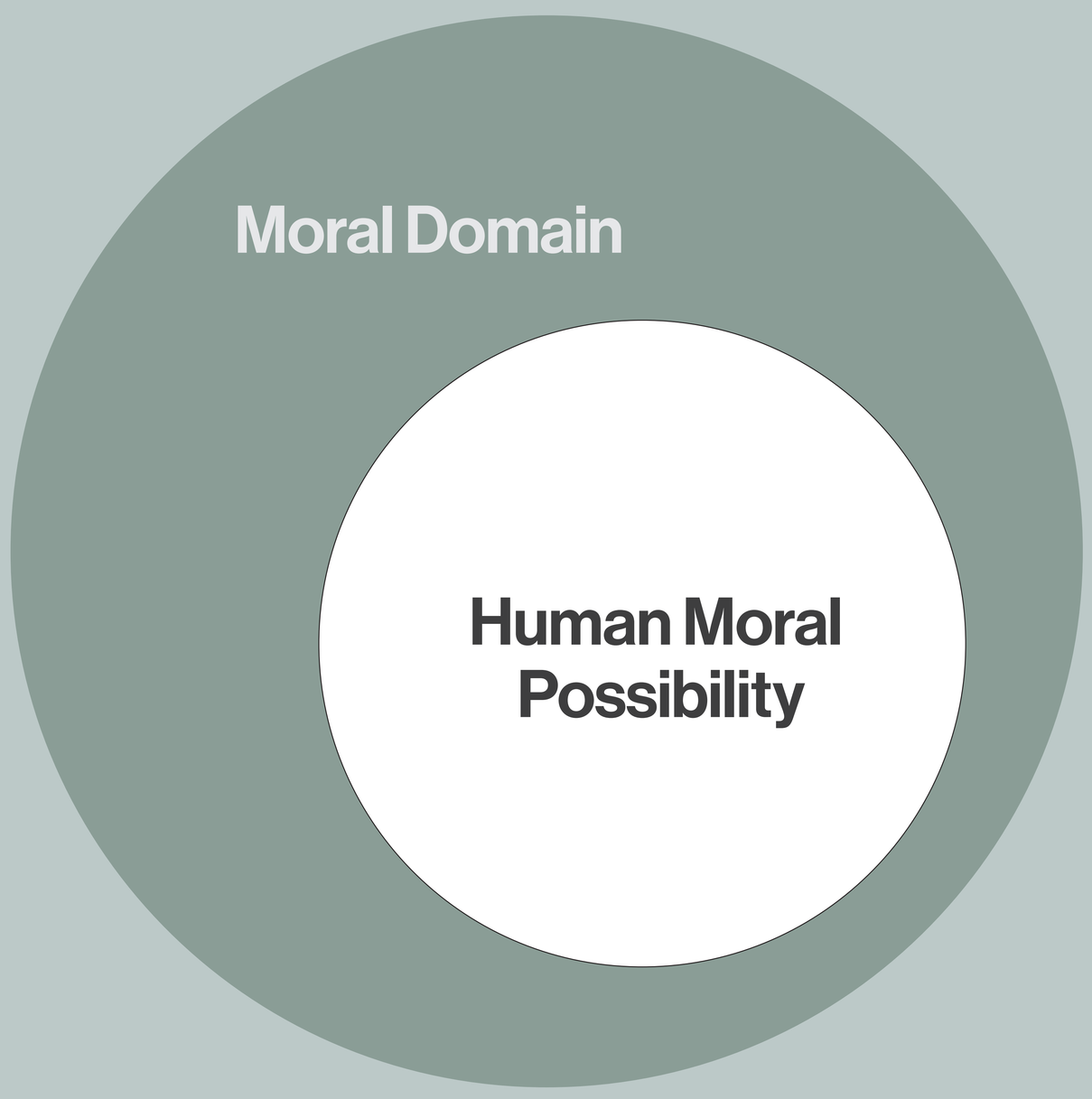
Jonathan Pengelly
Independent Researcher, Atarau

PhAI 2025, Amsterdam
October 23rd - 24th, 2025

Moral Cartography and Machine Ethics

Main Argument: Morality is broader and more diverse than its human manifestation. Machine ethics provides useful tools for systematically exploring and examining the full space of moral possibility.



The Moral Domain

The moral domain is abstractly concerned with rightness, goodness, and value and extends beyond human moral possibility. Configurations exist that are coherent and recognisably moral, yet unrealisable or unintelligible from a human standpoint given our limits.

Moral Cartography

Moral cartography is the systematic exploration of the full space of moral possibility—both within and beyond human limits. It maps both the terrain of the moral domain and the structural possibilities available to engage with it.

Machine Ethics

Machine ethics , by allowing us to view machines as potential moral subjects with their own position in moral space, provides three complementary cartographic instruments for exploring moral possibilities within and beyond the boundaries of human morality.

Cartographic Instruments

The View from Elsewhere

Adopting the hypothetical machine perspective provides critical distance from human morality, revealing its contingent features. This novel form of conceptual analysis requires the disciplined imagination to ask: How would X apply to an entity that doesn't share our limits?

Exploratory Simulation

Simulation allows us to explore what lies beyond human moral limits. Multi-agent frameworks let us construct agents with non-human characteristics. Their emergent strategies and behaviours provide conceptual raw material for identifying novel moral structures.

The Engineering Process

Building systems that replicate moral competencies transforms engineering into a philosophical instrument. Implementation demands precision, keeping us honest by revealing hidden complexity, operationalising assumptions and highlighting dependencies.

Why it matters

For moral philosophy: Moral cartography reveals contingent features of human morality, illuminating the boundaries that define it, unexplored possibilities within these boundaries, and possibilities beyond them.

For machine ethics: Opens a new exploratory research direction focused on philosophical discovery, complementing existing research areas.

For all: Recognising human morality as one region within a vast space of moral possibility cultivates epistemic humility about what we know of the moral domain and the limits of that knowledge.

Experiments

Using neuroevolutionary and multi-agent reinforcement learning techniques, I construct agents with non-human characteristics to explore what strategies and behavioral patterns emerge with the aim of identifying novel conceptual terrain.

Read the paper. View the code.