Moral Cartography and Machine Ethics

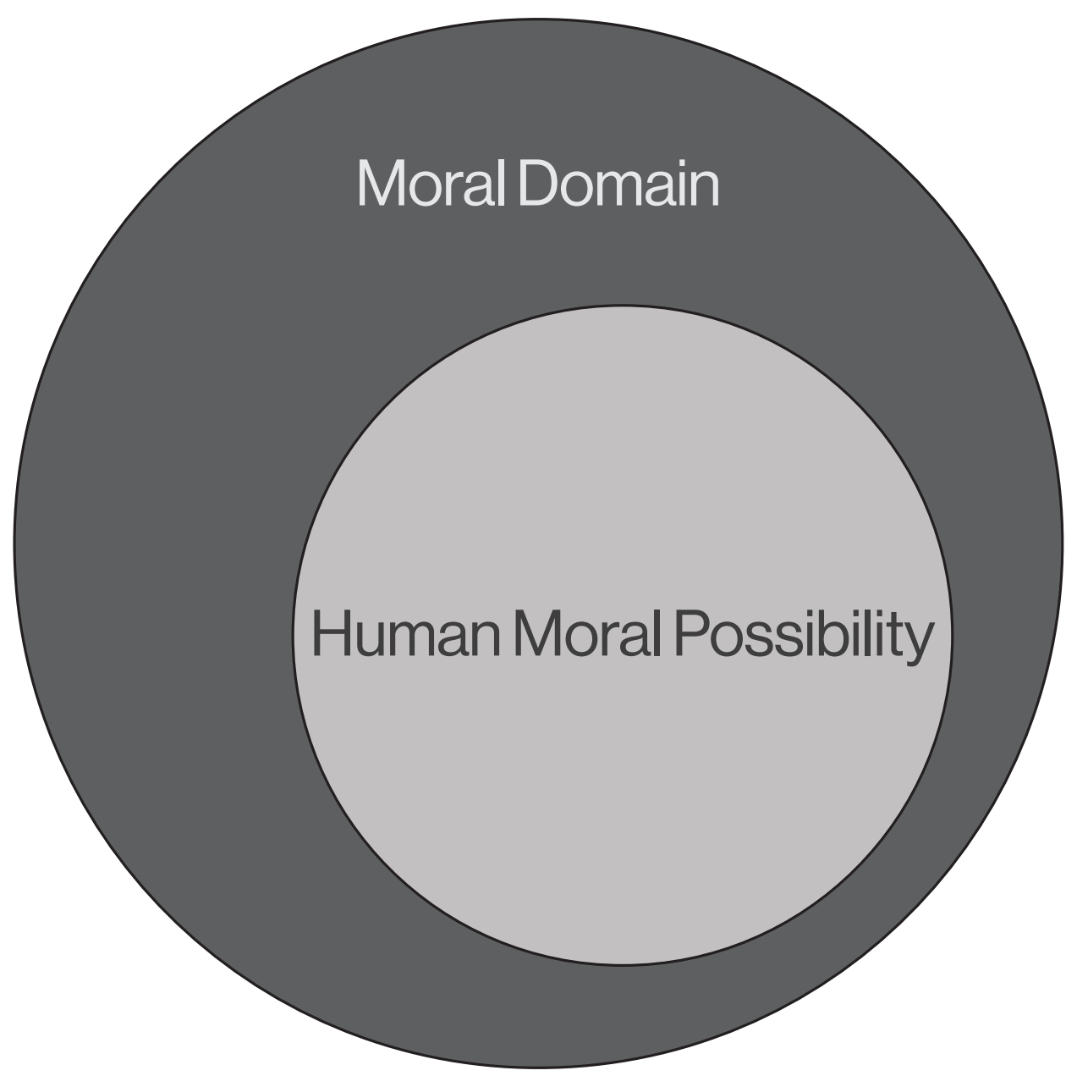
Jonathan Pengelly
Independent Researcher, Atarau

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

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. This content-oriented definition is independent of any particular manifestation.



The moral domain extends beyond human moral possibility. Configurations exist that are coherent in themselves, recognisably moral in nature, yet unrealisable or unintelligible from a human standpoint given our limits.

Concepts mediate both our understanding of the domain's boundaries and our engagement with it. While the domain is defined by its abstract features, concepts provide the interpretive frameworks that make moral concerns recognisable, intelligible and actionable.

Moral Cartography

Moral cartography is a proposed research program for the systematic exploration of the full space of moral possibility—both within and beyond human limits. It maps both the moral terrain and pathways through it (i.e. the domain regions as well as the structural possibilities to engage with them).

Building on Existing Work
Like moral anthropology [Flanagan 2016], which studies varieties of human moral possibility across cultures, moral cartography asks what forms morality can take. But where anthropology descriptively charts existing human territories, moral cartography ventures further, seeking to understand both the full extent of human moral possibility (realised and unrealised) and the terrain beyond these boundaries.

Terra Incognita
This requires new methods and shared frameworks. Drawing inspiration from conceptual engineering's emphasis on structured methodology and common vocabulary, moral cartography aims to establish systematic approaches for exploring how different structures make different regions of the moral domain intelligible and accessible.

The Cartographic Question
Rather than asking "What should X be?" moral cartography asks "What can X be?" It aims to explore how moral concepts can be interpreted and what different configurations are possible.

Machine Ethics

Machine ethics is "ethics for machines, for 'ethical machines', for machines as subjects, rather than for the human use of machines as objects" [Müller 2025].

An Alternative Standpoint
Machine ethics provides precisely the alternative perspective moral cartography requires. By viewing machines as potential moral subjects with their own position in moral space, it creates sufficient distance to examine human morality critically. This reveals how our moral concepts, frameworks, and practices are shaped not only by essential moral concerns but also by distinctly human limitations.

Cartographic Instruments
This alternative standpoint enables three distinct approaches for exploring moral possibilities within and beyond human boundaries, each detailed below: the view from elsewhere, exploratory simulation, and the engineering process itself.

These three instruments are complementary: the view from elsewhere identifies human boundaries, engineering forces explicit specification of what crosses them, and simulation systematically explores what lies beyond.

Cartographic Instruments

The View from Elsewhere

By adopting the hypothetical machine perspective, we gain a concrete standpoint sufficiently distant from human morality to reveal its contingent features.

A Thought Experiment
This novel form of conceptual analysis does not require any actual machines, nor any commitment to their potential existence, just the disciplined imagination to ask: How would X apply to an entity that doesn't share our human limitations? This perspectival approach yields concrete insights into both the boundaries of human moral concepts and the possibilities beyond them. [Pengelly 2023]

Example: Moral Aspiration
Unrealisable moral ideals have been shown to provide instrumental benefits for humans due to our cognitive and psychological limits. For entities without these constraints, such aspiration may serve no function, revealing it as a distinctly human scaffolding rather than fundamental to morality itself.

Example: Supererogation
Within human moralities, acts "beyond the call of duty" presuppose shared human limits that shape mutual expectations. Without such limits, the boundary between duty and supererogation becomes less clear-cut, showing how the concept depends on contingent features of the social dimension of human morality. [Pengelly 2025]

The Engineering Process

Building systems that replicate moral competencies transforms the engineering process itself into a philosophical instrument. Each implementation attempt has the potential to reveal hidden conceptual nuances and previously unrecognised dependencies.

Forced Precision
Engineering often demands unique philosophical rigour: moral competencies must be operationalised, dependencies made explicit, assumptions converted into code. Such work keeps philosophy honest [Dennett 2006]. Not only does it avoid imprecision, but it also reveals hidden complexity by showing the composite nature of concepts - some elements are essential, while others are contingent on human psychology or historical accident.

Learning from Failure
The failures and limitations encountered often prove particularly cartographically revealing. When implementations interpret situations in alien or absurd ways, we learn which aspects of our morality are contingent versus more general. The machine becomes a probe, simultaneously mapping human moral boundaries and identifying possibilities beyond them.

Beyond Self-Understanding
As Wallach and Allen noted, teaching machines right from wrong requires "attention to aspects of moral decision-making that people normally take for granted" [Wallach and Allen 2009]. But the cartographic potential extends further: we don't just learn about human morality but also consider what forms morality might take under different constraints.

Exploratory Simulation

Where the machine perspective reveals human moral limits, simulation explores what lies beyond them. Multi-agent learning frameworks enable us to construct agents with non-human characteristics (e.g. flexible identity boundaries, enhanced knowledge sharing and coordination, novel cognitive constraints). We can observe what strategies and behavioural patterns emerge through the application of various machine learning techniques.

Systematic Variation
The power lies not in replicating human morality but in systematic experimentation. We can define agents unconstrained by biological limitations, place them in environments with novel conditions, then discover what proves optimal or stable. Initial discoveries emerge as game-theoretic solutions that require moral interpretation. This provides us with the conceptual raw material for identifying unfamiliar moral structures.

Interpretive Challenges
Simulation results require careful analysis. Distinguishing genuinely novel behaviours from experimental artifacts is challenging but essential to exploratory simulation. Computational discoveries are not self-interpreting. They must be complemented by philosophical examination to yield meaningful conceptual insights. Nevertheless, simulation allows us to discover novel patterns that give us insight into regions of the moral domain that remain unintelligible within human frameworks.

Why it matters?

For moral philosophy: Reveals which features of human morality are contingent on human limits versus more general features of moral engagement, 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, which complements existing threads focused on implementation and potential integration within existing moral frameworks.

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

Experiments

These experiments use simulation as a cartographic instrument within an exploratory prototyping process.

Using neuroevolutionary/MARL techniques and LLMs, I construct agents with non-human characteristics (fluid identities, fragmented deliberation, constrained memory) to explore what strategies, situations, and reasoning models emerge and develop new conceptual tools where existing human concepts prove inadequate.

Each experiment aims to identify novel conceptual terrain, both within and beyond human possibility, seeking patterns and behaviours that, when interpreted through a moral lens, are not adequately captured by existing human moral concepts.

**Read the paper.
View the code.**