Moral Cartography and Machine Ethics

JONATHAN PENGELLY

Building on work showing morality to be more than its human manifestation, this paper introduces 'moral cartography' as a research program for the systematic exploration of this wider space of moral possibility. The central argument is that machine ethics provides an important set of instruments for this cartographic project. This instrumental function is illustrated through three key applications: (1) using the hypothetical 'machine perspective' to reveal the contingencies of human moral concepts; (2) employing simulation to both model and analyse novel conceptual possibilities; and (3) sourcing insights from the actual engineering of 'ethical machines' by identifying and analysing hidden dependencies and failure points when implementing human moral concepts. For moral philosophy, the cartographic approach offers a method for investigating the structure of morality beyond its human manifestation, cultivating both a more nuanced understanding of the scope of the moral domain and a crucial epistemic humility. For machine ethics, this approach represents a new exploratory line of research in which the emphasis is on philosophical discovery, rather than implementation and integration. By mapping the terrain of the moral domain, we illuminate not only what machine morality could be, but also where human morality stands in relation to it.

Additional Key Words and Phrases: moral cartography, machine ethics, abstract moral domain, machine morality, human limits

In 1610, Galileo Galilei published *Sidereus Nuncius*, reporting observations that would reshape our understanding of the cosmos. His telescope revealed mountains and valleys on the Moon, undermining Aristotelian notions of heavenly bodies as perfect and unchanging spheres. He also discovered moons orbiting Jupiter, which directly contradicted geocentric models placing the Earth at the center of all celestial motion. These observations did not merely add to existing astronomical knowledge; they challenged fundamental assumptions concerning the structure of reality and humanity's place within it.

This paper argues that machine ethics can serve an analogous function for moral philosophy. Just as the telescope revealed celestial features beyond the reach of human vision, I contend that by examining how machines engage with moral concepts and practices, we can uncover possibilities beyond the boundaries of human moral experience. Building on previous work showing that morality is broader and more diverse than its human manifestation[Pengelly 2023], this paper shifts the focus from examining specific moral concepts through the lens of a hypothetical machine perspective to the systematic exploration of moral possibilities beyond human limits.

Recognising this space of moral possibility is only the beginning; we also need methods to explore, map, and learn about it. This paper proposes *moral cartography* as a framework for such systematic exploration. I argue that machine ethics provides essential tools for this cartographic project. Three approaches will be discussed: the hypothetical machine perspective, simulation, and the engineering of 'ethical machines'. Each illustrates a different way machine ethics can serve moral cartography by helping us better understand the full breadth and diversity of the moral domain.

This is not merely an abstract theoretical exercise; moral cartography offers clear value for both moral philosophy and machine ethics. For moral philosophy, it helps us recognise the contingency of human morality, differentiating it from what is fundamental to morality itself. This opens up alternative ways of engaging with the moral domain while also casting human morality in a new light. For machine ethics, moral cartography represents a new exploratory line of research that emphasises philosophical discovery, complementing existing work on developing 'ethical machines' and their possible integration into the moral world. Perhaps most importantly, it cultivates an epistemic humility by allowing us to better understand the space of moral possibility, our own place within it, and the limits of our knowledge about it.

MORAL CARTOGRAPHY

The *History of Cartography* defines maps as "graphic representations that facilitate a spatial understanding of things, concepts, conditions, processes, or events in the human world."[Harley and Woodward 1987] This captures something essential about maps: they don't just record what exists; they also make it comprehensible. Moral cartography follows this same logic, mapping the moral domain to help us grasp the full scope and diversity of moral possibility.

But what exactly are we mapping? Following Pengelly's content-oriented definition, the moral domain is concerned with notions of rightness, goodness and distinctive forms of value.[Pengelly 2023] This definition is particularly useful because it remains independent of the human manifestation of morality. Indeed, human morality represents just one region within this broader domain, its boundaries shaped by our particular biological, cognitive, psychological, and social limitations.

Why then map territories we may never inhabit? The value of moral cartography lies in better understanding what MacIntyre refers to as the "space of moral possibility."[MacIntyre 2013] While MacIntyre focused on human moral possibilities, I expand this space beyond its possible human instantiation. Moral cartography thus promises to help us both recognise the boundaries of human morality more clearly and imagine possibilities that lie beyond them.

This distinction becomes clearer through a concrete example. Rather than asking 'What is agency?' or 'What should agency be?', moral cartography asks 'What forms of agency are conceivable?', 'How might they differ from each other?', and 'How do they impact other moral concepts?'. More generally, it represents a form of moral inquiry that asks not 'What should X be?' but rather 'What can X be?', both in terms of how X can be interpreted and what different X's are possible. This shift from analysis and prescription to exploration shows why moral cartography opens new territories for investigation.

Of course, not all cartography serves the same purpose. Some maps aim for greater accuracy or novel representations of familiar territories. Moral cartography embraces a different ethos, prioritising exploration over precision. It values the process of mapping new conceptual terrain. The goal is not to create definitive maps, but to chart previously unmapped territories, accepting that initial efforts will be provisional and incomplete.

TERRA INCOGNITA

In *The Geography of Morals*, philosophical anthropologist Owen Flanagan employs a cartographic metaphor strikingly similar to ours, in examining the "varieties of moral possibility"[Flanagan 2016] across human cultures. His work reveals how different worldviews and social arrangements generate distinct moral frameworks. These range from Confucian role ethics to Western individualism, from Buddhist teachings that view righteous anger as inherently destructive to Western traditions that see it as morally justified, to varied cultural understandings of justice and virtue. This diversity empirically demonstrates that morality is not monolithic but rather a complex moral topography shaped by history, culture, and circumstance.

Flanagan's cartographic approach illuminates how differently human societies conceptualise moral questions and the varied practices and frameworks we employ to navigate ethical problems. As he observes, "Read some literature, history, anthropology, or sociology and the varieties of moral possibility open up."[Flanagan 2016] His work shows how the particularities of history, gender, ethnicity, and socioeconomic status influence what Charles Taylor called our "orientation in moral space."[Taylor 1989] Indeed, Flanagan's empirically based mapping reveals not only what human morality is, but what it has been and might yet become.

Yet Flanagan's cartography, valuable as it is, charts only human territories. This focus is natural—a philosophical anthropologist necessarily focuses on human moral diversity. Moreover, as Flanagan rightly argues, our blindness to existing moral possibilities already narrows our perspective and constrains our thinking about alternatives. Significant work remains to better understand the domain of human morality.

But cartography need not stop there. Just as regional geography evolved into global exploration, and terrestrial mapping to cosmic cartography, moral cartography can extend its gaze beyond human boundaries. The shift from mapping human moral diversity to exploring the broader moral domain represents a fundamental shift in scale and ambition—from charting local territories to recognizing our position within a vast universe of moral possibilities.

This expansion becomes even more compelling when we consider Flanagan's observation about unexplored territories within human morality itself. "Some of the varieties of moral possibility have been tried and tested, but perhaps not in our time," he writes. "Most possibilities are unexplored, not yet conceived, and thus are terra incognita."[Flanagan 2016] If vast unexplored territories exist even within human moral possibility, how much vaster, then, must be the territories beyond?

This raises a provocative question: How might moral cartography help us understand both the unexplored regions of human morality and the territories that lie beyond human limits? The promise of moral cartography lies precisely in this dual vision. By mapping territories we may never inhabit, we gain perspective on the ground where we stand. By exploring moral possibilities beyond human constraints, we better understand which of our moral concepts are parochial, which touch on something deeper, and where the seeds for moral innovation might lie. The exploratory cartographer doesn't just return from uncharted territories with exotic tales, but also a transformed understanding of home.

TOWARDS A CARTOGRAPHER'S TOOLKIT

When Polynesian navigators set out across the Pacific, they didn't sail blindly into open ocean. They had knowledge of star paths and techniques for reading swells, currents, and the flight of birds. Centuries later, when European explorers arrived, they too came prepared with compass and sextant, logbooks and charts. Despite sailing into the unknown, both groups understood that discovery requires more than courage. It demands instruments for navigation, methods for recording findings, and crucially, ways to preserve knowledge for those who would follow in their wake.

Exploring the moral domain is no different. It also requires its own toolkit if we are to transform blind wandering into systematic discovery. Without shared methods and frameworks, we risk producing scattered insights that cannot build upon each other. In the recent development of conceptual engineering as a philosophical discipline, I see a model for how such a toolkit might emerge. It shows how a unifying research framework can transform informal practices into systematic methods that promote collaboration and knowledge sharing.

Conceptual engineering describes what philosophers have always done: analyse concepts, identify weaknesses, propose alternatives, examine implications. Yet the formalization of conceptual engineering as a field has transformed this scattered practice into something more powerful. It provides shared terminology for precise communication across research programs, develops methodological frameworks for different kinds of interventions, and creates standards for evaluation. Most importantly, it has created a community of practice where researchers can build systematically on each other's work.

Moral cartography requires a similar evolution. We need to see it not as individual exploration, but as a collective enterprise. At this initial stage, however, we face a fundamental challenge: it is difficult to specify constructively what tools we might need before we have begun the exploration in earnest. We cannot design the toolkit in advance as it

must emerge from the practice of exploration itself. No doubt we will learn what instruments we need as we discover the questions the territory poses.

Recognising that we need systematic methods while being unable to fully specify them in advance is itself valuable. It acknowledges both the necessity of developing shared frameworks and the inherent difficulty of doing so at the outset of a new intellectual enterprise. What we can do is identify examples like conceptual engineering to look to for guidance.

This is where machine ethics becomes not merely useful but essential. I contend that it provides precisely the kinds of philosophical and computational tools that moral cartography requires. Through these tools, we might explore moral territories we cannot directly inhabit, test configurations we cannot embody, and trace implications we cannot intuitively grasp. In the next section, we examine how these tools might serve the cartographic project.

MACHINE ETHICS AND CARTOGRAPHIC INSTRUMENTS

Machine ethics, as Vincent Müller defines it, is "ethics for machines, for 'ethical machines', for machines as subjects, rather than for the human use of machines as objects."[Müller 2025] How might it contribute to moral cartography? Machine ethics offers precisely the alternative standpoint that moral cartography needs: it views machines as potential moral subjects with their own position in moral space and critically examines how they might navigate ethically relevant questions. Three approaches stand out as particularly promising, each providing distinct methods for exploring the space of moral possibilities that extend beyond the boundaries of human morality.

The View from Elsewhere

The most immediately accessible cartographic use of machine ethics is perspectival. By adopting the hypothetical machine perspective, we gain a plausible standpoint sufficiently distant from human morality to critically examine its distinctive features. This approach 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 approach has already yielded concrete conceptual insights. By examining moral concepts through the machine perspective, Pengelly highlights idiosyncrasies of human morality that map poorly onto machines.[Pengelly 2023, 2025] For instance, unrealisable moral aspiration provides instrumental benefits due to human cognitive and psychological limits that don't translate to machines. Similarly, supererogation, acts that in some way go 'beyond the call of duty', presupposes shared human limits that shape mutual expectations. Furthermore, this approach identifies opportunities for novel conceptual interpretations, such as utilising the flexibility of artificial identity to explore forms of individual and collective agency unrealisable for humans.

Importantly, this perspectival approach requires only conceptual analysis, avoids contentious claims about actual moral machines, yet yields concrete insights into both the boundaries of human moral concepts and the possibilities that exist beyond them. Like an anthropologist using cross-cultural comparison to identify what's universal versus culturally specific, the moral cartographer can use the hypothetical machine perspective to distinguish contingent elements of human morality from fundamental features of morality itself.

Simulation

Where the machine perspective excels in revealing the limits of human morality, simulation allows us to explore what lies beyond them. Modern multi-agent reinforcement learning (MARL) frameworks are one particularly promising

approach for moral discovery through simulation. They enable us to construct agents with non-human characteristics (e.g. flexible identity boundaries, enhanced knowledge sharing) and observe what strategies, behaviors, and coordination patterns emerge under these alternative conditions.

The power of simulation lies not in replicating human morality but in systematic variation. We can define agents, both individual and collective, that are unconstrained by biological limitations, place them in environments with novel physical or social laws, then use machine learning techniques to discover what practices prove optimal or stable. Many of these discoveries will initially appear as mere game-theoretic solutions or coordination strategies rather than recognisably moral behaviors. Yet this is precisely their cartographic value: they provide conceptual raw material such as patterns of cooperation, resource distribution, identity management, that can then be transplanted and reinterpreted within moral frameworks.

This highlights a fundamental interpretive challenge with MARL: discoveries are not self-interpreting. The difficulty lies in distinguishing novel, morally-interesting behaviours from mere artifacts of the experimental setup. For example, what appears as an innovative form of cooperation might simply reflect a quirk of the reward function or agent model. This hermeneutic problem is by no means insurmountable. Indeed, it is part of the challenge of the moral discovery process. However, it does mean simulation must be complemented by careful analysis to ensure promising computational results do in fact represent meaningful moral insights.

Despite these interpretive challenges, this approach is already yielding empirical discoveries. Current exploratory work using DeepMind's Melting Pot framework[Agapiou et al. 2023; Leibo et al. 2021] has already delivered intriguing results that hint at the possibilities. For example, agents with fluid identity boundaries develop forms of cooperation that transcend human notions of individual and collective. When they can share experiences directly, validation practices emerge that assume perfect information rather than managing uncertainty. These aren't corruptions of human concepts and practices, but valid alternative solutions to coordination problems under different constraints.

While this work remains preliminary, it demonstrates how experimental work with computational frameworks can help us begin mapping regions of moral space that would otherwise have remained purely speculative. Indeed, the simulations need not be sophisticated; like early telescopes, even simple instruments can reveal previously invisible features of the landscape.

The Engineering Process

A third cartographic approach emerges from the actual engineering of systems designed to replicate moral competencies. The engineering process itself becomes a philosophical instrument, with each implementation attempt potentially revealing hidden conceptual nuances and previously unrecognised dependencies.

Wallach and Allen recognised this potential in 2009, noting that teaching machines right from wrong requires "attention to aspects of moral decision-making that people normally take for granted," calling it "an exercise in self-understanding."[Wallach and Allen 2009] But the cartographic potential extends beyond self-understanding. When we attempt to engineer moral competencies, we don't just learn about human morality but also consider what forms morality might take under different constraints.

The engineering process forces a unique kind of philosophical rigour. Every moral competency must be operationalized, every dependency made explicit, every assumption converted into code or architecture. As Dennett observed, this keeps us honest.[Dennett 2006] Yet beyond avoiding imprecision, engineering also reveals hidden complexity. Implementation often reveals the composite nature of concepts—some elements prove essential, others prove contingent

on human psychology or historical accident. Ultimately, each engineering decision maps another feature of moral terrain.

Perhaps most valuably, the failures and limitations encountered while building these machines prove cartographically revealing. When these implementations interpret situations in ways that seem alien or absurd to us, we often learn which aspects of our morality are contingent and which are more general. The machine becomes not just an artifact but a probe, simultaneously mapping the boundaries of human morality and revealing possibilities beyond them.

These three approaches represent complementary ways in which machine ethics can serve moral cartography. Each reveals different features of the moral domain: the machine perspective provides the critical distance needed to recognise the boundaries of human morality; simulation explores new territories within the space of moral possibility; and engineering reveals hidden complexities through implementation. Together, they provide the initial tools required to transform moral cartography from speculative idea into actual method.

Crucially, these instruments are available now. As the cartographic project develops, we can expect both our tools and our maps to grow more sophisticated. But machine ethics allows us to get started with the task of mapping the moral domain, however imprecisely. Indeed, this is how such speculative research must begin: not with perfect instruments but with the tools at hand. Machine ethics provides these tools so that we might venture beyond the familiar shores of human morality and begin charting new spaces of moral possibility.

THE VALUE OF MORAL CARTOGRAPHY

Why venture beyond the familiar territories of human morality? Why not concentrate instead on concrete ethical challenges deserving our attention? The answer lies not in abandoning human concerns, but in enriching our understanding by mapping the full expanse of the moral domain—and discovering our place within it.

The parallel with abstract mathematics is instructive. When first developed in the 19th century, non-Euclidean geometry appeared as a mere intellectual curiosity, fascinating but useless speculation about impossible spaces. Yet, ultimately, it proved essential for Einstein's general relativity, revealing that the universe itself is non-Euclidean. Mapping moral possibilities beyond human experience might seem equally impractical—just a theoretical curiosity with no immediate application. Perhaps examining these seemingly impossible moral structures will likewise illuminate hidden features of our moral world.

Beyond this speculative value, however, moral cartography does offer tangible benefits. For moral philosophy, the hypothetical machine perspective proves valuable for conceptual analysis. It represents a plausible standpoint sufficiently distant from human morality to critically examine its distinctive features. Indeed, this framing alone raises compelling meta-ethical questions about the nature and scope of the moral domain itself.

Most significantly, moral cartography cultivates epistemic humility without lapsing into relativism or moral scepticism. Just as astronomy advanced by abandoning geocentric theories, moral philosophy benefits from recognising that human morality, while central to us, occupies just one region in a vast space of moral possibility. This shift in perspective doesn't diminish the importance of moral values, nor does it require abandoning existing moral theory. Rather, it provides a clearer view of how human limits have shaped our particular moral world. Such clarity may even promote moral innovation by expanding our awareness of unrealised possibilities within human morality itself.

For machine ethics, adopting a cartographic perspective fundamentally reframes core questions in the field. Instead of asking how to integrate machines into human moral frameworks, we might consider how human and machine moralities could function as distinct regions within the broader moral domain. This reframing offers new possibilities

for alignment research as well. A focus on identifying areas of productive coexistence may prove more fruitful than pursuing convergence irrespective of fundamental differences.

More practically, moral cartography represents a new research direction for machine ethics, one focused on philosophical discovery over implementation and theoretical integration. It gives researchers room to explore and experiment, while avoiding some of the charged debates around moral standing and conformance to human moral standards. Sometimes the most practical path forward is to step back and map the territory, so that we might have the tools to decide how and where best to build.

The value of moral cartography ultimately lies in its potential to shift our thinking toward a more exploratory approach to the moral domain. This doesn't mean abandoning human morality or embracing relativism. A good cartographer can map territories she would never choose to inhabit. But by understanding the true scope of moral possibility, we gain something precious: the ability to see our own moral commitments clearly, to engage thoughtfully with radical moral difference, and to navigate wisely as our world grows stranger and more complex.

CONCLUSION

This paper has proposed moral cartography as a systematic framework for exploring the space of moral possibilities. Just as the telescope revealed celestial features invisible to the naked eye, machine ethics offers instruments to begin mapping these new moral territories. These tools, the hypothetical machine perspective, simulation, and the engineering process itself, are already revealing the vastness of the moral domain and our particular position within it.

The ultimate promise of moral cartography is not to replace human morality but to understand it better, so that we might see clearly, perhaps for the first time, where we stand in the larger landscape of moral possibility. In an age where we rapidly create new forms of intelligence, such understanding becomes not merely intellectually satisfying but practically essential. After all, the territories we map today may well be the spaces we all inhabit tomorrow.

REFERENCES

John P. Agapiou, Alexander Sasha Vezhnevets, Edgar A. Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, DJ Strouse, Michael B. Johanson, Sukhdeep Singh, Julia Haas, Igor Mordatch, Dean Mobbs, and Joel Z. Leibo. 2023. Melting Pot 2.0. arXiv:2211.13746 [cs.MA] https://arxiv.org/abs/2211.13746

Daniel Dennett. 2006. Computers as Prostheses for the Imagination. Invited talk presented at the International Computers and Philosophy Conference. May 3, 2006 Laval, France.

Owen Flanagan (Ed.). 2016. *The Geography of Morals: Varieties of Moral Possibility.* Oxford University Press.

J.B. Harley and David Woodward (Eds.). 1987. *The History of Cartography: Volume 1.* The University of Chicago Press.

Joel Z. Leibo, Edgar Dué nez Guzmán, Alexander Sasha Vezhnevets, John P. Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charles Beattie, Igor Mordatch, and Thore Graepel. 2021. Scalable Evaluation of Multi-Agent Reinforcement Learning with Melting Pot. *International conference on machine learning.* https://doi.org/10.48550/arXiv.2107.06857

Alasdair MacIntyre. 2013. On Having Survived the Academic Moral Philosophy of the Twentieth Century. In *What Happened in and to Moral Philosophy in the Twentieth Century?*, Fran O'Rourke (Ed.). University of Notre Dame Press.

Vincent C. Müller. 2025. Ethics of Artificial Intelligence and Robotics. In *The Stanford Encyclopedia of Philosophy* (Fall 2025 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

Jonathan Pengelly. 2023. *The possibilities of machine morality.* Ph. D. Dissertation. Victoria University of Wellington, Wellington, New Zealand. https://doi.org/10.26686/wgtn.23539932

Jonathan Pengelly. 2025. Machine Supererogation and Deontic Bias. In *Governing the Future: Digitalization, Artificial Intelligence, Dataism*, Henning Glaser and Pindar Wong (Eds.). CRC Press, 96–107.

Charles Taylor. 1989. *Sources of the Self: The Making of the Modern Identity.* Harvard University Press.

Wendell Wallach and Colin Allen. 2009. *Moral Machines: Teaching Robots Right From Wrong.* Oxford University Press.

Received July 7, 2025