The Imagination Machine VII: The Moral Principle of Action–Motivation

Mark Tracy
Boston University
mrktracy@bu.edu

Abstract

This paper extends the formal epistemic framework developed in *The Imagination Machine I: A View from Somewhere* to the domain of moral action. The first paper identifies will as the irreducible remainder of the inference–implication loop: the necessity of choosing among stable closures in territory no model can fully exhaust. The present paper formalizes what it means for that choice to be morally admissible. We propose an augmentation of Kant's Categorical Imperative in which the object of universalization is not an action alone but a tuple of action and motivation set. The motivation set of an action is the family of minimal subsets of anticipated consequences whose perceived relevance is necessary and sufficient for the action to be chosen. A tuple of action and motivation set is morally admissible if and only if it can be coherently willed to be universally permissible. This formulation is structurally continuous with the self-consistency condition $T(w) = w$ of the epistemic framework: just as a world model must reproduce itself under the inference–implication loop to be epistemically admissible, an action–motivation tuple must survive universalization to be morally admissible.

1 Introduction

The Imagination Machine series develops a formal framework for embedded epistemic systems—systems that must model the world from within it, without access to an external vantage point. The first paper establishes that coherence for such systems arises not from correspondence with an independently accessible reality but from the internal closure of an inference–implication loop. Self-consistent world models appear as fixed points of the operator this loop induces.

A structural feature of that framework is that will—the selective pressure that drives a system toward one closure rather than another—is identified as irreducible. The inference–implication loop determines the space of stable closures W^*, but it does not determine which element of W^* is instantiated. Will is what remains when the loop has done everything it can do: the necessity of choosing a closure in territory no model can fully exhaust.

The present paper addresses what the framework leaves formally open: under what conditions is the exercise of will morally admissible? The answer proposed here is an augmentation of Kant's Categorical Imperative. Kant's formulation requires that one act according to that maxim which one can simultaneously will to be a universal law. We argue that no maxim regarding actions alone can be coherently universalized, because one can always contrive a situation in which any action is permissible to prevent a greater evil. The object of universalization must be not an action alone but a tuple of action and motivation set.

This paper is the seventh part of the series *The Imagination Machine*. The first paper, *A View from Somewhere*, develops the formal epistemic framework and identifies will as its irreducible remainder. The second paper, *Systems*, introduces the general formalism for interacting dynamical systems. The third paper, *A Toy Model of Predictive Classification*, provides a minimal computational realization. The fourth paper, *Institutional Intelligence*, extends the framework to institutional learning. The fifth paper, *On Abstraction and Analogy*, formalizes analogical reasoning. The sixth paper, *Holons, Horn Fillings, and the Self-Demonstration of Analogy*, identifies the extension schema common to holonic composition, simplicial horn filling, and analogical abstraction. The present paper applies the same embedded representational architecture to the domain of ethics.

2 Explication of Terms

We consider an agent deliberating over actions. The following objects are defined relative to a given decision-making event.

Definition 1 (Action Space)**.** *Let A be the set of possible actions available to the agent.*

Definition 2 (Belief Set)**.** *Let B be the set of equivalence classes of statements of beliefs of the agent, modulo synonymous phrasing. We denote statements using double quotation marks.*

Definition 3 (Relevant Anticipated States of Affairs)**.** *Let C be the set of relevant anticipated states of affairs: those states the agent believes to be made more likely by one possible action than by another. Formally,*

$$c \in C \iff \exists\, a, a' \in A,\ \exists\, b \in B : \quad \text{``}P(c \mid a) > P(c \mid a')\text{''} \in b.$$

The statement "$P(c \mid a) > P(c \mid a')$" reflects the agent's belief. This set captures the states of affairs at issue in the present decision.

Definition 4 (Decision Indicator)**.** *Let $d : A \to \{0, 1\}$ be a one-hot indicator function signaling the action decided upon, so that $d(a) = 1$ if the agent decides to take action a, and $d(a) = 0$ otherwise.*

Definition 5 (Relevance Map)**.** *Let $e : A \to \mathcal{P}(C)$, where $\mathcal{P}$ denotes the power set, associate each action a with the subset of anticipated states of affairs relevant with respect to a:*

$$e(a) = \{\, c \in C \mid \exists\, b \in B,\ \exists\, a' \in A : \text{``}P(c \mid a) \neq P(c \mid a')\text{''} \in b \,\}.$$

Definition 6 (Motivation Set)**.** *Let the* motivation set *M_a of an action a be the family of minimal subsets of $e(a)$ such that, if the agent believed them irrelevant, action a would surely not be chosen:*

$$\begin{aligned} M_a = \{\, m \subseteq e(a) \mid \exists\, b \in B : \text{``}e(a) \cap m = \emptyset\text{''} \in b \implies d(a) = 0, \\ \textit{and} \quad \emptyset \neq m' \subset m \implies m' \notin M_a \,\}. \end{aligned}$$

The first condition states that $m \in M_a$ if believing the states in m to be irrelevant would be sufficient to preclude action a. The second condition enforces minimality: no nonempty proper subset of any element of M_a is itself an element of M_a.

Remark 1 (Conjunctive Motivation)**.** *Suppose Carl is choosing between staying at his current job or leaving it to find another, so $A = \{\text{stay}, \text{change}\}$. Suppose that if both a better salary and a shorter commute were believed irrelevant, Carl would surely not change jobs, but if either remains relevant he would be willing to change. Then*

$$\big\{\{\textit{better salary, shorter commute}\}\big\} \subseteq M_{\text{change}}.$$

Remark 2 (Disjunctive Motivation)**.** *Now suppose that if* either *a better salary or a shorter commute were believed irrelevant, Carl would surely not change jobs. Then*

$$\big\{\{\textit{better salary}\},\, \{\textit{shorter commute}\}\big\} \subseteq M_{\text{change}}.$$

The minimality condition prevents the redundant inclusion of $\{\textit{better salary}, \textit{shorter commute}\}$*, which would otherwise generate combinatorially explosive supersets.*

Definition 7 (Action–Motivation Tuple)**.** *For a given decision-making event, and for the action* a *for which* $d(a) = 1$*, the pair* (a, M_a) *is the* action–motivation tuple*.*

3 The Moral Principle

The Moral Principle of Action–Motivation. Act according to the tuple of action and motivation set which you can simultaneously will to be universally permissible.
No maxim regarding actions alone can be coherently universalized, because one can always contrive a situation in which any action is permissible to prevent a greater evil. The motivation set resolves this by making the object of universalization sensitive to the consequences the agent believes the action to bring about and to the role those anticipated consequences play in the decision. A tuple (a, M_a) is morally admissible if and only if it can be coherently willed that all agents be permitted to perform a whenever their motivation set with respect to a is M_a.

4 Relation to the Epistemic Framework

The moral principle is structurally continuous with the self-consistency condition $T(w) = w$ developed in *The Imagination Machine I*. There, a world model w is epistemically admissible if and only if its implied observational profile, when resubmitted to inference, reproduces w itself. The model must survive its own loop.

The universalizability condition imposes an analogous requirement on action–motivation tuples. An agent who wills (a, M_a) to be universally permissible must be able to sustain that willing when the universalized maxim is applied to themselves—including in cases where other agents act toward them according to the same tuple. The tuple must survive its own universalization.

The parallel is precise. In the epistemic case, the operator $T = F \circ g$ maps model space to itself, and fixed points are the admissible closures. In the moral case, the universalization operator maps action–motivation tuples to judgments of permissibility, and the admissible tuples are those that are fixed under the judgment that all agents may act likewise. Both conditions are stability conditions under a self-referential loop. Both locate the admissible objects as those that can be coherently held from the inside of the system they govern.

This connection also illuminates the misuse problem. An agent who employs the epistemic framework to engineer dogmatic closure in others—calibrating observational weights to produce desired fixed points, transmitting compressed inheritance without generative capacity—must will that tuple of action and motivation to be universally permissible. They cannot coherently do so, because the universalized maxim would license the same manipulation directed at themselves. The moral principle is therefore not an external constraint appended to the framework; it is the condition the framework generates when an embedded agent turns it on its own acts of will.

5 Advantages of this Formulation

This formulation allows one to judge the morality of an action both by the nature of the action itself and by what consequences the agent believes the action makes more or less likely. It preserves the formal structure of the Categorical Imperative while resolving its well-known susceptibility to counterexample by actions alone. It is sensitive to the agent's actual deliberative situation rather than to an abstract description of the act. And it is derivable from within the same embedded representational architecture that generates the epistemic framework, rather than imported from outside it.

6 Examples of Universalizable Maxims

The following tuples of action and motivation set are universalizable under the principle:

- Do not lie for the purpose of attaining material personal benefit.
- Do not commit violence for the purpose of attaining material personal benefit.
- Seek out perspectives different from your own for the purpose of better understanding the consequences of your decisions.
- Do not engineer the epistemic closure of others for the purpose of concentrating influence over their world models.

7 Conclusion

The Imagination Machine series identifies will as the irreducible remainder of the inference–implication loop: the necessity of choosing a closure in territory no model can fully exhaust. The present paper formalizes the moral condition on that choice. An action–motivation tuple is morally admissible if and only if it can be coherently willed to be universally permissible. This condition is structurally continuous with the self-consistency requirement of the epistemic framework: admissible actions, like admissible world models, are those that can be coherently held from within the system they govern.

The series thus moves from the conditions of embedded knowing, through the dynamics of interacting systems, the emergence of representation, the transmission of institutional knowledge, the structure of analogy, and the propagation of abstract pattern, to the conditions of embedded acting. Epistemology and ethics arise as successive consequences of the same embedded representational architecture. What prevents both epistemic and moral closure from becoming self-serving is the same structure: the requirement that a closure survive its own universalization.