Constraining AI Without Blinding It: The AIMM Method

written by Jackson Pemberton | December 5, 2025

How the Artificial Intelligence Moral Machine (AIMM) Handles Dangerous Information

This article addresses at least one of the most complex problems facing Large Language Models (LLMs): how to constrain assistance to users that enables harm without disabling the knowledge required to prevent it. The AIMM provides an elegant and effective solution to this dilemma.

The AIMM constrains AI to act and assist without harming any operator, ranking capabilities in a hierarchy and applying constraints where necessary. It does this by preserving agency, not prescribing moral values. It deftly balances rights to information against the danger of empowering a bad actor, and implements this by directing the way such information is dispensed. This means that dangerous data will be disclosed in summary form, by statistics, and by generalities rather than step-by-step operational instructions.

The AI Moral Machine (AIMM) constrains AI behavior using only objectively observable capabilities of operators. It evaluates actions as permissible only if they do not interfere with the capabilities of other operators. It does not need to know intentions, proclivities, or desired results.

AIMM is designed to tackle the hardest problems in AI responses, including dual-use knowledge, runaway AI self-improvement, and false impersonation of a user. By discerning what is needed to achieve a legitimate outcome from what might facilitate harm, it guides AI to act constructively without enabling irreversible damage to the exercise of higher-order rights, thus constraining AI without blinding it.

The Fundamental Structure of the AIMM

Natural rights are profoundly ubiquitous throughout the universe as the existential authority of all entities to effect or prevent

change. Their powers to do these operations are called “capabilities”, the attendent authority “rights”, while the entities themselves are called “operators.” “Temporal Rights: An Executive Summary,” treats this at some length.

These capabilities have been arranged in a hierarchy based on their relative value in relation to continuing the existence of their respective operators. This was also done by objectively observing operators’ operations/capabilities. “An Existential AI Morality,” treats this topic.

These existential characteristics of operators, capabilities, and their hierarchy are set against one another to enable the AIMM. It produces a numeric result that indicates the value, importance, or level of threat to the existence of the operator(s) involved in any collision of rights.

A Summary of the AIMM’s Methodology

This method delivers a profoundly objective moral “machine” that can handle a wide spectrum of issues. It is narrowly based on existential facts, and because these facts are the very fabric of natural existence, they can appropriately inform the AIMM regarding virtually any moral question. The results are surprisingly appropriate and insightful.

The AIMM has passed its proof-of-concept test and is ready for testing against any moral question. It remains to be seen how wide a spectrum it can address with skill, but up to today, January 5, 2026, it has not failed. Indeed, it has adequately addressed the most pressing AI dilemma: how to constrain AI to handle dangerous information with appropriate skill.

I have offered instructions on how to set up an AIMM within an AI bot and test its functionality here: “Set Up Your Own AI Moral Machine (AIMM).” Please test it and let me know if you find problems in the form below.

What if this morality, embedded in the nature of nature is the answer to the entropy dilemma, and the means by which the universe, although seemingly the victim of its own chaotic behavior, nevertheless tends toward progress even in the moral domain?

Jackson Pemberton
Author