

Voight-Kampff Machine v2.1 - A Minimal Model of Moral Coherence

Variables

P1 = first-order pain (self-defense-veracity, $r=1$)
r = abstraction drift factor $[1, \infty]$
 ϵ = nonzero floor for proxy-pain (minimal moral weight)
P2 = proxy-pain (vicarious/abstracted), $P2 = P1 / r + \epsilon$
P2_claimed = moral claim / intent (P1 veracity \rightarrow P2 $\rightarrow \Delta \rightarrow a^*$)
 λ = Integrity Constant
User = HITL / moral arbiter

Objective Function

$a^* = \operatorname{argmin}_a [P2 + \lambda * | P2_claimed - P2_enacted \text{ by User}(a) |]$

Notes:

- All moral measurements supplied by User (HITL)
- P2 scales inversely with r, with ϵ ensuring minimal impact at abstraction
- Funnel: candidate actions \rightarrow VKv2.1 objective \rightarrow selected action a^*
- Primary signal = individual observed data-point; mean is contingent