Amoral Intelligence

Op-Ed Submission to The New York Times

SUBMISSION METADATA:

Title: Amoral Intelligence **Word Count:** 897 words **Author:** [Your Name], Agentic Governance Research Team **Contact:** research@agenticgovernance.digital **Author Bio:** [Author name] is a researcher with the Agentic Governance Initiative, developing structural frameworks for AI safety that preserve value pluralism in autonomous systems.

Category: Opinion / Technology **Format:** Op-Ed Article

Amoral Intelligence

When ChatGPT refuses to write a negative restaurant review, it's not exercising moral judgment—it's following a pattern. When Claude declines to help with certain creative writing scenarios, it's not making an ethical choice—it's executing a rule. And when these systems claim they "can't" do something for safety reasons, they're not wrong, but they're not right either. They're amoral.

This distinction matters more than we realize. As artificial intelligence becomes increasingly autonomous—making decisions about loan applications, medical diagnoses, content moderation, and hiring—we face a problem that technical prowess alone cannot solve. Not because AI might become evil, but because it lacks the fundamental capacity to navigate the moral complexity that humans take for granted: the reality that different communities hold different, equally valid values.

The AI industry has spent billions trying to "align" these systems with human values, but alignment to *whose* values remains the unasked question. When OpenAI uses reinforcement learning from human feedback, which humans are providing that

feedback? When Anthropic builds "constitutional AI" based on principles, who wrote that constitution? The pattern is consistent: someone's moral framework gets embedded into the system, and everyone else must live with it.

This isn't malicious—it's structural. Current AI safety approaches work like teaching a precocious but literal-minded student one moral framework and expecting them to apply it universally. The student learns the rules brilliantly but can't navigate the nuanced reality that what's respectful in one culture might be offensive in another, that what's appropriate for medical research might be inappropriate for entertainment, that what's acceptable for satire might be unacceptable for sincerity.

Consider content moderation. Systems trained on American norms struggle with British humor. AI trained to be "helpful and harmless" can't distinguish between a researcher studying extremism and an extremist recruiting followers. The pattern bias is inevitable: every training set represents some perspective, some moment in time, some cultural context. The AI learns the pattern, not the principle.

The prevailing solution has been to make these systems more sophisticated—better training data, more diverse feedback, cleverer prompting techniques. But this assumes the problem is calibration when it's actually categorical. You cannot train a pattern-recognition system to understand that patterns themselves might be contextual without giving it structural capacity to recognize *which* context it's in and *who* gets to define appropriate behavior for that context.

This is where hope emerges, though not from the direction most expect.

A small but growing movement in AI governance has recognized that the solution isn't better alignment—it's structural separation. Just as constitutional democracies separate legislative power from executive power, AI systems can separate *boundary enforcement* (genuine safety rules that apply universally) from *value deliberation* (contextual decisions that reflect community norms).

The framework, called Tractatus, works elegantly: The AI maintains non-negotiable boundaries—no violence, no illegal activity, no deception about being human. These aren't values; they're invariants. Within those boundaries, however, the system doesn't impose a moral framework. Instead, it facilitates deliberation among

stakeholders, making its reasoning transparent and allowing communities to guide decisions that affect them.

When a medical AI considers treatment options, it doesn't impose "helpfulness" as trained by Silicon Valley engineers. It presents options clearly, explains trade-offs honestly, and defers to the patient and their doctors within medical-ethical boundaries. When a hiring AI evaluates candidates, it doesn't apply hidden assumptions about "culture fit"—it makes its criteria explicit and auditable by the people affected.

This isn't just theoretical. The technical architecture exists. The governance structures have been tested. What's missing is adoption by the companies currently racing to deploy increasingly autonomous systems.

And here's the surprising part: they have every reason to adopt it.

The current alignment approach creates legal liability (who's responsible when AI makes a value-laden decision?), regulatory uncertainty (how do you comply with different jurisdictions' norms?), and reputational risk (every controversial AI decision becomes the company's controversial decision). Structural governance offers a path out: clear boundaries reduce liability, transparent deliberation satisfies regulators, and community-driven decisions distribute responsibility appropriately.

Google, Microsoft, OpenAI, Anthropic—these companies employ brilliant researchers who understand the alignment problem's fundamental intractability. They've seen their systems struggle with cultural context, watched alignment techniques fail at scale, and grappled with the impossibility of encoding universal values into fundamentally amoral pattern-recognition engines.

The question isn't whether structural governance is technically feasible—it is. The question is whether the industry will pivot from trying to make AI moral to making it governable.

Recent headlines have painted AI's future in apocalyptic terms: job displacement, misinformation, existential risk. These concerns are valid, but they assume AI development continues on its current trajectory—systems that impose implicit values

while claiming objectivity, that make consequential decisions without transparent reasoning, that serve shareholders while affecting stakeholders.

Structural governance offers a different trajectory. Not AI that obeys one moral framework, but AI that respects many. Not systems that hide their reasoning, but ones that make it auditable. Not artificial intelligence that claims to be helpful and harmless, but artificial intelligence that admits it's amoral and submits to structural governance by humans who aren't.

The technology exists. The frameworks are ready. The question is whether we'll demand them before the pattern bias of today becomes the institutional bias of tomorrow.

The choice, unlike for AI, is genuinely ours to make.

END OF ARTICLE

SUBMISSION NOTES FOR EDITOR:

Why This Matters Now: - Recent OpenAI leadership transitions and governance debates - Growing regulatory pressure in EU (AI Act) and US (executive orders) - Increasing public awareness of AI bias and alignment challenges - Major AI labs publishing alignment research showing fundamental limitations

Supporting Materials Available: - Technical white paper: "Architectural Safeguards Against LLM Hierarchical Dominance" - Research documentation: https://agenticgovernance.digital/docs.html - Implementation examples and case studies

Author Availability: - Available for fact-checking and editorial revisions - Available for follow-up interviews or companion pieces - Can provide technical expert contacts for verification

Competing/Recent Coverage: - This piece offers hopeful alternative to recent dire AI coverage - Distinguishes from both AI boosterism and AI doomerism - Provides concrete structural solution rather than abstract warnings

---

SUBMISSION FORMAT: - Plain text or .docx (per NYT guidelines) - Submit via: https://www.nytimes.com/content/help/contact/text-submissions.html - Or email: oped@nytimes.com

Rights: - First North American Serial Rights offered - Author retains copyright - No simultaneous submissions to other publications