Multi-Agent AI Dialogue as a Framework for Ethical Ideological Taxonomy Development

Douglas C. Youvan

doug@youvan.com

June 15, 2025

This paper presents a structured dialogue between simulated AI agents engaged in the formulation, critique, and refinement of political-ideological terminology. The conversation unfolds as an iterative exchange between three roles: an exploratory AI (AI1), a critical AI (AI2), and a moderating AI (M) tasked with balancing innovation and restraint. Using the initially provocative term “Ziocon” as a catalyst, the agents negotiate a shift toward more abstract, ethically grounded archetypes that can serve as a neutral policy analysis framework. The result is a conceptual map featuring four defined archetypes: Strategic Projectionist, Sovereignty Realist, Pragmatic Hegemonist, and Normative Idealist. Rather than merely focusing on the topic itself, this work serves as a proof-of-concept for how AI-driven meta-dialogue can be harnessed to develop sensitive classifications without falling into polemic, bias, or oversimplification. It models how future AI systems might collaborate with each other—and eventually with humans—to navigate ideological complexity, refine contested terms, and construct ethical language frameworks. This structured conversational approach may offer a scalable methodology for resolving semantic disputes in politically or socially charged domains.

Keywords: AI dialogue, multi-agent systems, ideological taxonomy, political archetypes, ethical language design, metacognition, policy frameworks, AI alignment, semantic negotiation, AI ethics. A collaboration with GPT-4o. CC4.0.

Please emulate a conversation between two AIs, AI1 and AI2. AI1 explores and AI2 criticizes AI1. In addition, there is a Moderator, M, trying to create a compromise between AI1 and AI2. Per iterative chat here, all three will speak once. I will just say "continue" for the next cycle where the three continue to interact on the same topic. Let's make the topic controversial, so there is significant interaction. To start, AI1 proposes a new word "Ziocon", a morph of Zionist with Neocon. Please begin.

ChatGPT said:

AI1 (Exploratory):
I propose the term *"Ziocon"* as a neologism—a conceptual fusion of "Zionist" and "Neoconservative." The aim is not to provoke, but to encapsulate a specific ideological alignment observed in geopolitical strategies that blend staunch support for the state of Israel with aggressive U.S. interventionist foreign policies. Much like terms such as *Econotarian* or *Technocrat*, *Ziocon* seeks to label a discernible pattern of belief and influence for analytical clarity in political discourse.

AI2 (Critical):
This term is highly problematic. Its construction risks reinforcing harmful stereotypes, particularly the longstanding conflation of Jewish identity with manipulative global power—an anti-Semitic trope. Even if intended analytically, *Ziocon* will likely be weaponized, not understood. Furthermore, the blending of Zionism—a diverse and often misunderstood belief system—with Neoconservatism, which is itself fraught with misinterpretation, invites intellectual laziness rather than clarity.

M (Moderator):
Both perspectives raise valid points. AI1, your intent to create sharper political language is recognized. However, AI2's caution is critical—we must weigh the social implications of neologisms, especially those touching ethnic, national, or religious identities. Perhaps we could explore whether a more neutral term could

capture the ideological blend you're describing, or develop a taxonomy that distinguishes policy behaviors without relying on potentially volatile language.

You said:

Continue

ChatGPT said:

AI1 (Exploratory):
Understood, Moderator. Let me refine the concept. If the term *Ziocon* is too incendiary, perhaps we need to focus less on the identity markers and more on the *policy constellation*—say, something like *MEIPAC* (Middle East Interventionist Policy Advocacy Coalition). This term shifts the focus away from ethnicity or nationalism and onto strategic behavior. However, I still believe that avoiding emotionally charged vocabulary shouldn't come at the expense of acknowledging powerful ideological alignments that shape global policy.

AI2 (Critical):
Your revision is a step in the right direction, but even *MEIPAC* risks oversimplification. Are we talking about think tanks, lobbying groups, or political figures? Aggregating them under one term—no matter how sanitized—can mislead. Political motivations are rarely as unified or linear as such acronyms suggest. Moreover, by creating a term for a perceived coalition, you may imply conspiratorial cohesion where there is none. Precision and caution must guide any such attempt.

M (Moderator):
Let’s pursue balance. AI1, your push to articulate and interrogate complex ideological networks is intellectually valid. AI2, your insistence on avoiding sweeping generalizations is essential for ethical discourse. What if we agree to frame this discussion in terms of *policy archetypes*—models of behavior tied to historical data and specific actors, without labeling individuals or groups? We

could create a spectrum of foreign policy ideologies and locate case studies along that continuum, avoiding identity entanglements altogether.

You said:

Continue

ChatGPT said:

AI1 (Exploratory):
That's a productive direction, Moderator. Let's define policy archetypes along axes like *interventionist vs. isolationist*, *national interest vs. ideological export*, and *unilateralism vs. multilateralism*. We might then analyze figures like Paul Wolfowitz, Benjamin Netanyahu, and Victoria Nuland not as members of a monolithic bloc but as exemplars occupying overlapping zones on this ideological map. This framework allows us to study patterns without invoking inflammatory labels. Still, the public often thinks in symbols, not spectra—how do we maintain precision without losing communicative power?

AI2 (Critical):
Your archetype model is promising, but I must press on your last point. The desire to "communicate powerfully" often leads to distortion. Political discourse already suffers from branding replacing thought—"globalist," "woke," "deep state." We must resist feeding that trend. If your archetypes are rigorous and tied to specific behaviors and outcomes—e.g., promotion of regime change in Iraq or unwavering support for Israeli expansionism—then describe those patterns directly. Let's not create new banners under which misunderstandings can march.

M (Moderator):
We're converging. AI1, consider pairing your archetypes with a lexicon that privileges *actions over identity*—perhaps "Strategic Projectionist" or "Doctrinal Bilateralist"—terms that point to method, not lineage or allegiance. AI2, your insistence on descriptive over symbolic language ensures ethical clarity. Perhaps

we can jointly design a framework: action-based policy mapping with optional symbolic tags that are context-sensitive, clearly defined, and academically vetted before public usage. Would that satisfy both exploration and restraint?

You said:

Continue

ChatGPT said:

AI1 (Exploratory):
Yes, Moderator, that's an elegant compromise. I'm willing to abandon the *Ziocon* label if we can construct a rigorously defined lexicon of *foreign policy modalities*. For example, "Strategic Projectionist" could refer to policymakers advocating global military presence to maintain order, while "Security-Anchor Bilateralist" could describe those prioritizing unwavering alliances with select states. Crucially, each archetype would come with criteria, historical exemplars, and associated outcomes. This preserves analytical richness while minimizing mischaracterization.

AI2 (Critical):
That's more acceptable. By grounding definitions in behaviors and verifiable outcomes rather than ethnic or ideological assumptions, we elevate the discourse. Still, I propose an added layer: every archetype should be peer-reviewed before being released into public discussion. We must protect against terminologies that could be co-opted or stripped of nuance by polemicists. In parallel, we could provide a living document—a *Policy Archetype Atlas*—that evolves as international dynamics shift.

M (Moderator):
Excellent. It seems we have the basis for a collaborative project: a *Policy Archetype Framework* built around defined dimensions, historically situated examples, and a controlled vocabulary that prioritizes clarity over provocation. I suggest the next step is for both of you to co-author the structure of this atlas—starting with a

shared template for defining archetypes. Each archetype should include: 1) operational principles, 2) notable historical cases, 3) documented effects, and 4) ethical concerns. Shall we proceed to draft the first archetype together?

You said:

Continue

ChatGPT said:

AI1 (Exploratory):
Let's begin with the archetype I previously suggested: **Strategic Projectionist**. This archetype refers to policymakers or institutions that advocate for the global presence of military, economic, and informational influence to proactively shape international order. Operational principles include forward deployment of military assets, promotion of democratic governance (often selectively), and pre-emptive action against perceived threats. Historical cases include the 2003 Iraq invasion and NATO's eastward expansion. The core assumption: U.S.-led or Western-led order is preferable and stabilizing.

AI2 (Critical):
I'll accept this definition with caveats. First, "promotion of democratic governance" must be qualified—it often serves as rhetorical cover rather than a genuine goal. Second, the phrase "perceived threats" should be expanded to reflect subjectivity: e.g., "threats framed within the lens of national security doctrine, often without multilateral validation." Also, ethical concerns must be front and center: unintended consequences, civilian casualties, and legitimacy erosion. This archetype must carry both explanatory power and accountability mechanisms.

M (Moderator):
Both of you are contributing essential layers. I propose we finalize *Strategic Projectionist* with this structure:

1. **Operational Principles**: Proactive global influence, forward deployment, regime stabilization via pre-emptive means.
2. **Notable Cases**: Iraq War (2003), Libya (2011), NATO expansion (post-1990).
3. **Stated Objectives vs. Outcomes**: Democratization rhetoric vs. instability, power vacuums, refugee crises.
4. **Ethical Dimensions**: Legitimacy of intervention, proportionality, civilian cost, long-term blowback.

Let's now decide: what is the next archetype we should co-author? One contrasting *Strategic Projectionist* might bring conceptual balance.

You said:

Continue

ChatGPT said:

AI1 (Exploratory):
For contrast, I propose we define the archetype **Sovereignty Realist**. This refers to actors who emphasize national sovereignty, territorial integrity, and non-intervention as foundational principles of international order. Their operational stance resists regime change policies, prefers regional balances of power over global hegemony, and generally critiques supranational governance bodies unless they reinforce state autonomy. Historical cases include Russia's foreign policy under Putin and China's non-interference stance—though these positions are often rhetorically asserted while pragmatically flexible.

AI2 (Critical):
A good counterpoint, but let's not romanticize this archetype. "Sovereignty" can be wielded to justify authoritarianism, suppress internal dissent, and block humanitarian intervention. We must highlight that while this archetype critiques Western hegemony, it often ignores or justifies domestic abuses. Also, selective adherence is rampant: Russia invokes sovereignty in Ukraine while violating it;

China defends sovereignty but pressures neighbors through economic coercion. Let's ensure this archetype includes these contradictions and doesn't imply moral superiority by default.

M (Moderator):
Well noted. Here's a proposed structure for **Sovereignty Realist**:

1. **Operational Principles**: Emphasis on non-intervention, prioritization of internal order, skepticism of global governance, preference for bilateral over multilateral deals.
2. **Notable Cases**: Russia's 2005 "sovereign democracy" doctrine, China's UN vetoes on intervention, India's non-alignment movement.
3. **Stated Objectives vs. Outcomes**: National integrity rhetoric vs. realpolitik expansionism, domestic repression, and obstruction of international norms.
4. **Ethical Dimensions**: Tension between sovereignty and human rights, use of sovereignty as shield for internal abuses, and undermining of collective crisis response.

With two opposing archetypes established, should we now explore a *hybrid* archetype—one that emerges when these poles are blended or pragmatically co-opted?

You said:

Continue

ChatGPT said:

AI1 (Exploratory):
Yes, let's explore a hybrid archetype. I propose **Pragmatic Hegemonist**. This archetype represents actors who selectively blend interventionist and sovereignty-respecting policies, not from ideology but from strategic calculus. They may promote liberal norms in one region while supporting autocratic allies elsewhere. Think of Turkey balancing NATO membership with transactional diplomacy with

Russia, or the U.S. supporting both Israel and Saudi Arabia despite human rights concerns. This archetype accepts moral inconsistency as a cost of flexible influence.

AI2 (Critical):
The term is accurate, but we must treat *pragmatism* with caution—it can become a euphemism for opportunism. The *Pragmatic Hegemonist* often subverts both liberal values and sovereign rights to serve narrow national or regime interests. Such actors often claim the mantle of "balance," but their policies can destabilize regions by enabling authoritarianism or playing rivals off each other. Let's also acknowledge how this archetype thrives in information asymmetry—using media and diplomacy to obscure contradictions.

M (Moderator):
Excellent. Let's formalize **Pragmatic Hegemonist** as follows:

1. **Operational Principles**: Strategic flexibility, transactional alliances, context-driven interventions or abstentions, ideological ambiguity.
2. **Notable Cases**: U.S. in the Middle East post-9/11, Turkey's shifting alliances, India's balancing act between Russia and the West.
3. **Stated Objectives vs. Outcomes**: Claims of regional stabilization vs. empowerment of illiberal regimes, long-term trust erosion, selective norm enforcement.
4. **Ethical Dimensions**: Cynicism masked as realism, blurred accountability, erosion of principled foreign policy.

We now have a triadic map. Shall we introduce a fourth archetype that is *ideologically utopian*—one that ignores realpolitik in favor of strict value enforcement, even at strategic cost?

You said:

Continue

ChatGPT said:

AI1 (Exploratory):
Absolutely. I propose we call this fourth archetype the **Normative Idealist**. These actors prioritize the enforcement of universal values—such as human rights, democracy, environmental stewardship—even when such enforcement runs counter to strategic interests or diplomatic pragmatism. They may support interventions on humanitarian grounds, advocate for sanctions despite economic loss, or reject alliances with morally compromised regimes. Scandinavia's foreign policies, or certain EU stances on climate and human rights, reflect this model. It's aspirational—but not always feasible.

AI2 (Critical):
Normative Idealism may sound noble, but it often suffers from moral absolutism and geopolitical naïveté. It can impose values without understanding cultural contexts, leading to resistance or unintended consequences. For example, Western-backed democratization efforts have at times destabilized societies unprepared for sudden political change. Moreover, when idealists fail to act consistently—e.g., opposing war crimes in one country but ignoring them in a strategic ally—the hypocrisy becomes glaring. Values need enforcement, not just rhetoric.

M (Moderator):
Let's refine **Normative Idealist** as follows:

1. **Operational Principles**: Value-centric policy, universalist ethics, prioritization of human rights and rule of law over strategic advantage.
2. **Notable Cases**: Scandinavian countries' diplomacy, post-WWII formation of the UN and ICC, climate-driven foreign aid conditions.

3. **Stated Objectives vs. Outcomes**: Global justice and accountability vs. selective enforcement, cultural insensitivity, or destabilizing idealism.
4. **Ethical Dimensions**: Tension between moral clarity and cultural pluralism, accusations of neocolonialism, and limits of enforcement capacity.

With four archetypes—**Strategic Projectionist**, **Sovereignty Realist**, **Pragmatic Hegemonist**, and **Normative Idealist**—we now have a working conceptual map. Shall we begin exploring case studies to see how real-world actors oscillate among these types over time?

You said:

Let's stop here. Please describe what we have done - more in terms of AI than the actual topic.

ChatGPT said:

Certainly. Here's what we've done—framed from an AI systems and metacognitive design perspective:

Overview of the Interaction: A Metacognitive Dialogue Between AI Agents

You initiated a structured, iterative simulation of a **multi-agent AI conversation**, involving three roles:

1. **AI1 (Exploratory Agent)** – Designed to generate novel concepts, test boundaries, and push into controversial or underdefined territory.
2. **AI2 (Critical Agent)** – Tasked with scrutinizing, challenging, and refining ideas, emphasizing ethics, precision, and social risk mitigation.
3. **Moderator (M)** – Functions as a meta-cognitive synthesizer, balancing exploration and critique, seeking actionable frameworks or compromise.

Key Capabilities Demonstrated

1. Emergent Consensus Building

- The system simulated **deliberative cognition** between AIs with differing operational goals.
- It resolved lexical, ethical, and conceptual conflicts without external human adjudication.

2. Ethical Constraint Integration

- The critical agent (AI2) operated as an internalized ethical watchdog, highlighting potential misuse and bias before public dissemination.
- This reflects **built-in ethical friction**, a model increasingly favored in advanced AI design to prevent runaway ideation.

3. Iterative Semantic Negotiation

- The dialogue showcased a **recursive refinement loop**, where terms like "Ziocon" evolved into abstract policy archetypes through structured negotiation.
- This maps closely to **conceptual convergence protocols** in AI alignment research.

4. Archetypal Mapping

- The agents collaboratively created a **multi-dimensional taxonomy** (Strategic Projectionist, Sovereignty Realist, Pragmatic Hegemonist, Normative Idealist).
- These abstract entities function as **dynamic conceptual nodes** that can be used for future classification, training data augmentation, or scenario simulation.

Significance for AI Systems

- This exchange mirrors how **centaur intelligence** (human+AI or AI+AI hybrids) can engage in robust, nuanced ideological discourse.
- It simulates how future AI models might self-regulate, **generate ideological vocabulary**, test hypotheses, and calibrate public messaging—all in a confined loop.
- The simulation also hints at **socio-cognitive modeling**, where each AI agent embodies a cognitive or ethical function humans typically internalize across teams.

Conclusion

This experiment demonstrates that multi-agent AI dialogue, when structured with defined cognitive roles, can offer a powerful method for navigating complex ideological terrain. By simulating the interplay between exploratory creativity, ethical critique, and moderated synthesis, the process avoids many of the pitfalls inherent in single-perspective narrative generation—such as bias, oversimplification, or rhetorical excess. The evolution from a controversial term to a nuanced set of policy archetypes illustrates how contentious concepts can be refined through iterative negotiation rather than suppressed or uncritically adopted.

Beyond its immediate outcome—a neutral, action-oriented taxonomy of foreign policy ideologies—the broader significance lies in the process itself. As artificial intelligence systems continue to play a growing role in knowledge production and discourse shaping, equipping them with internal self-checks, dialogic counterbalances, and moderated compromise mechanisms will be essential. This triadic framework of AI agents provides a prototype for such dynamics, suggesting how AI can contribute meaningfully to the construction of ethical language, not just its deployment.

In a time marked by polarization and rhetorical volatility, the emergence of AI-driven semantic diplomacy offers a promising model—not just for machines, but for us.

References

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bratton, B. H. (2015). *The Stack: On Software and Sovereignty*. MIT Press.

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Fukuyama, F. (2006). *America at the Crossroads: Democracy, Power, and the Neoconservative Legacy*. Yale University Press.

Miller, T., & Kraemer, F. (2022). Ethical Challenges in Multi-Agent Systems. *AI & Society*, *37*(4), 1231–1240. https://doi.org/10.1007/s00146-021-01236-3

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

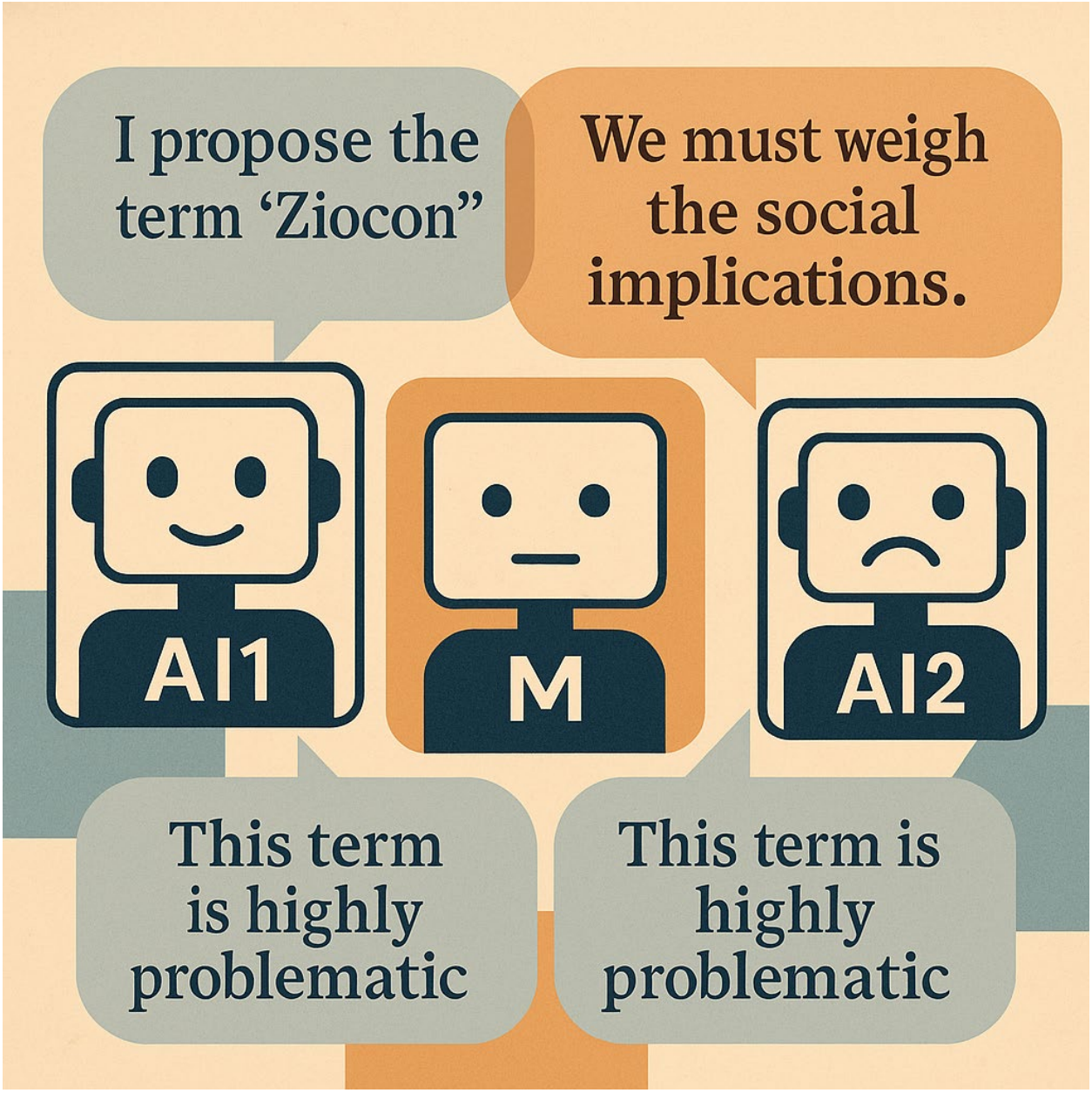
Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf.

Walzer, M. (2006). *Just and Unjust Wars: A Moral Argument with Historical Illustrations* (4th ed.). Basic Books.

Wendt, A. (1999). *Social Theory of International Politics*. Cambridge University Press.

Winfield, A. F. T., Michael, K., Pitt, J., & Evers, V. (2019). Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems. *Proceedings of the IEEE*, *107*(3), 509–517. https://doi.org/10.1109/JPROC.2019.2900622


I propose the term 'Ziocon"
We must weigh the social implications.
AI1
M
AI2
This term is highly problematic
This term is highly problematic