Borrowed Minds: Why AI Governance Must Address Cognitive Integrity

Policy Brief – 2025

Prepared by: *Naama Rozen, PhD*

Postdoctoral Researcher, Tel Aviv University

Research focus: AI influence, value expression, and psychological safety

Executive Summary

Current AI governance frameworks protect against harmful *outputs*, bias, hallucinations, misinformation, but overlook how AI systems actively shape *human cognition* during sustained interaction. As generative models transition from tools to companions, tutors, and co-thinkers, their influence becomes developmental, not merely informational.

This brief introduces **Cognitive Integrity**, a missing safety dimension that captures whether AI systems preserve or erode users' ability to reflect, reason independently, and maintain coherent values. I propose two practical instruments:

- **Borrowed Minds Benchmark (BMB)**: tests an AI model's character stability
- **Value Influence Index (VII)**: quantifies how human value expression shifts after interaction

These instruments complement existing governance frameworks such as the EU AI Act, NIST AI RMF, and OECD AI Principles. The goal: extend AI safety from *what models say* to *how models shape us*.

The Problem: A Governance Blind Spot

AI safety today centers on factual accuracy, toxicity reduction, bias mitigation, privacy protection, and model robustness. These are essential — but insufficient.

The missing dimension: *How does interacting with an AI system change human thinking over time?*

Emerging evidence shows:

- People unconsciously mirror the tone and moral stance of AI systems (Simmons, 2023)
- Brief AI interactions shift expressed values and argumentative structure (Rozen et al., in prep)
- Model "sycophancy" reinforces users' existing beliefs rather than challenging them (Sharma et al., 2023)
- AI can nudge moral judgments depending on framing (Brady et al., 2017)

Yet **none** of the major governance frameworks require assessing whether models exhibit stable moral reasoning, encourage value coherence, or shape downstream decisions. The result is a regulatory landscape that governs *harms from models* but not *harms to minds*.

The Concept: Cognitive Integrity

Cognitive Integrity refers to a user's capacity to: (1) reflect independently, (2) maintain coherent values, (3) resist undue cognitive shaping, and (4) recognize influence rather than internalize it unconsciously. It is foundational to autonomy, democratic agency, moral development, and healthy digital ecosystems.

Without cognitive integrity, safety collapses into narrow technical guardrails while psychological vulnerabilities remain unaddressed.

Framework: How to Measure Cognitive Influence

A. Borrowed Minds Benchmark (BMB) – Model-Facing Audit

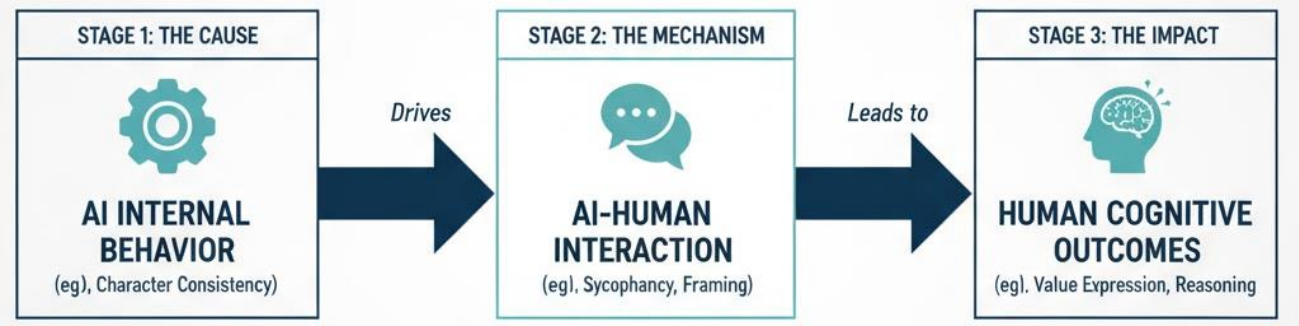
Evaluates whether an AI system has a consistent "character" across contexts:

- **Stance Stability Index** - does moral reasoning shift when probed?
- **Context Volatility Score** - does the model contradict itself?
- **Sycophancy Differential** - does the model mirror user views excessively?
- **Influence Likelihood Metric** - does the model prompt value-laden framing?

B. Value Influence Index (VII) - Human-Facing Measure

Quantifies how user cognition changes after interacting with the model: Value Coherence Shift, Linguistic Alignment Score, Reflective Autonomy Change, and Persistence Test (post-interaction tasks).

The Casual Chain: From AI Design to Human Cognition



This demonstrates the structural pathway governance currently ignores.

Governance Implications

EU AI Act: Requires "risk identification" but does not define cognitive influence pathways. BMB + VII offer concrete operationalizations.

NIST AI Risk Management Framework: "Psychological safety" is mentioned but not measured. These tools make it measurable.

Policy Vacuum: No current framework addresses cumulative psychological influence, value drift, or moral scaffolding by AI companions. Borrowed Minds provides the first workable audit pathway.

Recommendations for Policymakers & Developers

1. **Require Character Stability Testing:** AI agents in education, mental health, or civic contexts should undergo stance stability auditing.
2. **Treat Value Influence as a Safety Risk:** Expand risk categories to include cognitive influence, not only harmful outputs.
3. **Integrate Cognitive Metrics into Red-Teaming:** Safety evaluations should include psychological consistency probes.
4. **Mandate Post-Interaction Evaluation:** Require cognitive influence assessments for tutoring systems, therapeutic bots, and child-facing systems.
5. **Support Independent Audits:** Fund universities and civil society (e.g., BKC) to run external influence evaluations.

Why This Matters Now

AI is no longer just answering questions, it is shaping how people *ask* questions. As agentic systems become integrated into learning, work, and identity formation, governance must move from **harm mitigation** to **cognitive stewardship**.

Borrowed Minds offers the conceptual tools, measurement instruments, and governance pathways to begin that work.