Meaning Before Alignment

A consciousness-first account of AI ethics

The Science of Consciousness 2026, San Diego

Concurrent talk

Venkatesh H. Chembrolu

IKSMHA Center, Indian Institute of Technology Mandi

• Chembrolu et al., *Beyond Alignment: Representational Ethics and the Governance of Constructed Worlds*, AI and Ethics, 2026 DOI 10.1007/s43681-026-01355-1
• Prabhath et al. *Semantic Information Beyond Shannon: Meaning, Coherence, and Agency in Experiential Systems,* submitted (2026) https://www.researchsquare.com/article/rs-10284367/v1

Behavioral alignment and representational ethics

Behavioral alignment

Treats AI mainly as a producer of behavior and outputs

Primary test

Does the output satisfy a value, rule, preference, or safety constraint?

Representational ethics

Treats AI as a designer of personalized, configurable experiential realities

Primary test

How does technological mediation reorganize attention, interpretation, memory, and action?

The shift moves ethical evaluation from isolated outputs to the construction of worlds of meaning.

Representational ethics

Ethical evaluation of how designed representations construct, filter, and reorganize the spaces of meaning through which agents perceive, interpret, and act

Object of evaluation

The designed transformation and feedback process, not only the artifact's visible output

Ethical criterion

Whether the agent can understand, contest, and correct how meaning changes over time

Distinct contribution

Longitudinal representational drift becomes a specific ethical failure with design and governance levers

Semantic information supplies the account of meaning needed for this shift.

Consciousness as methodological ground

Meaning appears within experience before it becomes a metric.

The methodological claim

Conscious experience is the field in which representations become relevant, intelligible, and capable of guiding action.

What the argument does not require

No claim that present AI is conscious
No claim that computation produces consciousness
No transfer of moral agency from people to machines

Focus: how artificial systems reshape the experience of human agents

Shannon information and semantic information

Shannon information

Uncertainty reduction

Signal transmission

Statistical dependence

Disassociated from meaning,
Meaning can be derived

Semantic information

Aboutness and relevance

Action guidance

Stability under variation

Meaning is intrinsic
Explains why a distinction matters to an agent

$$SI = g(R, A)$$

R: representational organization
A: agentive alignment

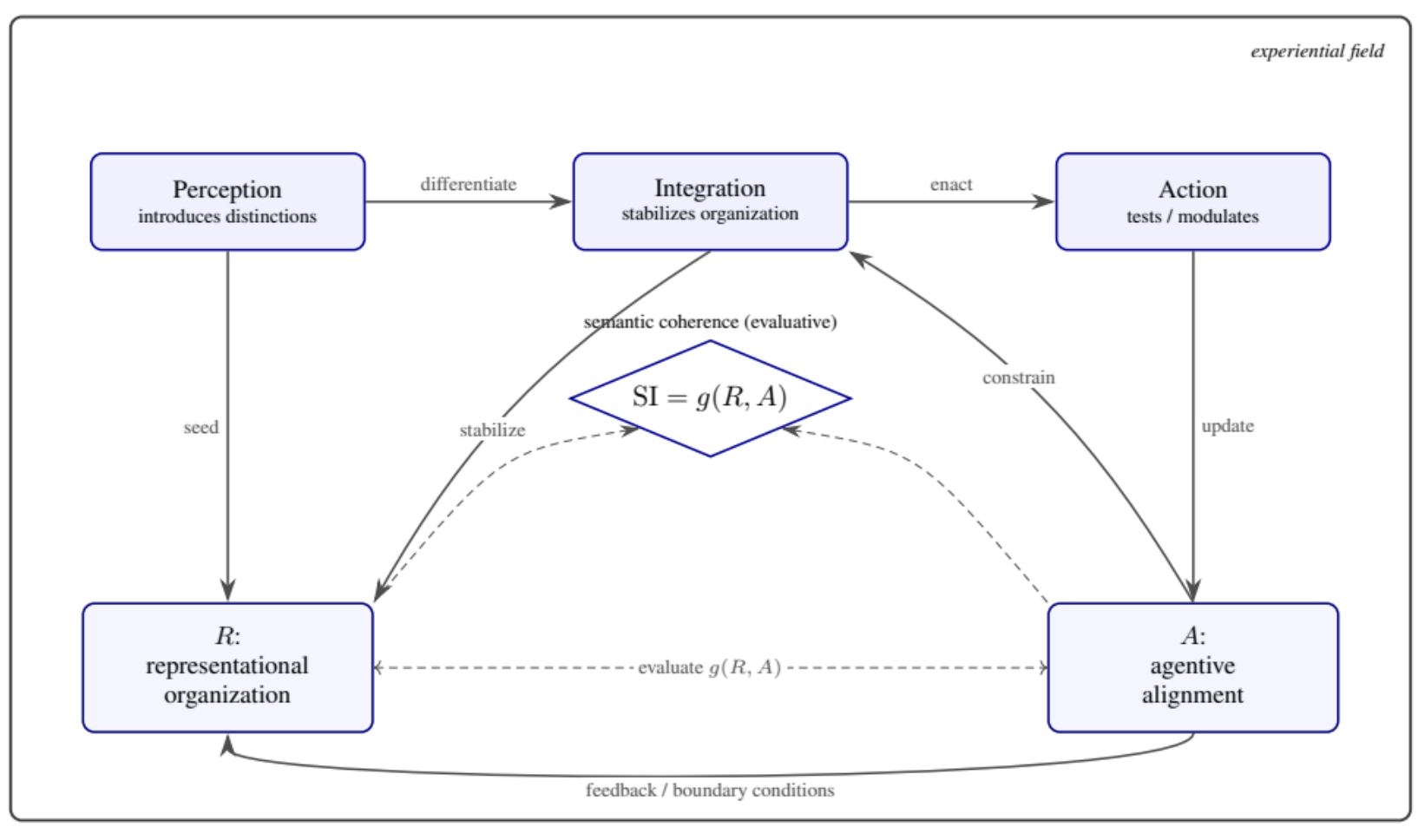
Meaning–Experience Coupling

Perception
introduces distinctions

Integration
stabilizes organization

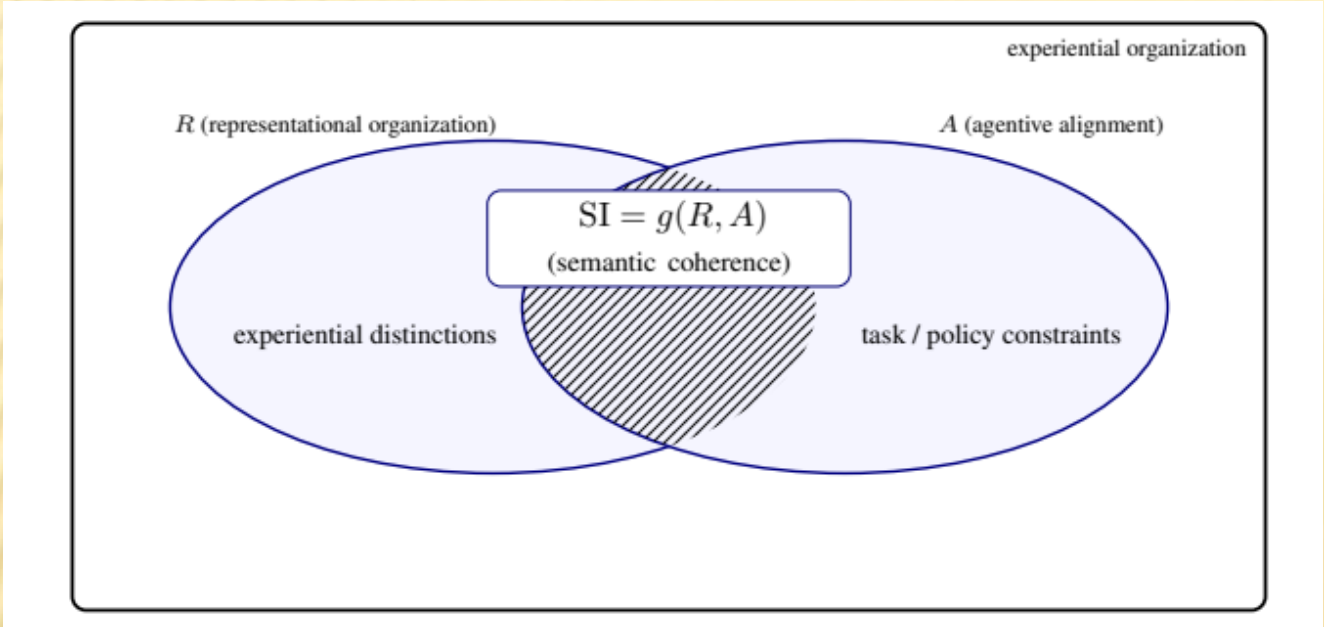
Action
tests and revises

SI evaluates coherence. It does not add another causal stage.



Source figure: Prabhath et al. *Semantic Information Beyond Shannon: Meaning, Coherence, and Agency in Experiential Systems*, submitted (2026)

Semantic information as coherence



$$SI(t) = g(R(t), A(t))$$

Graded — Meaning can be partial

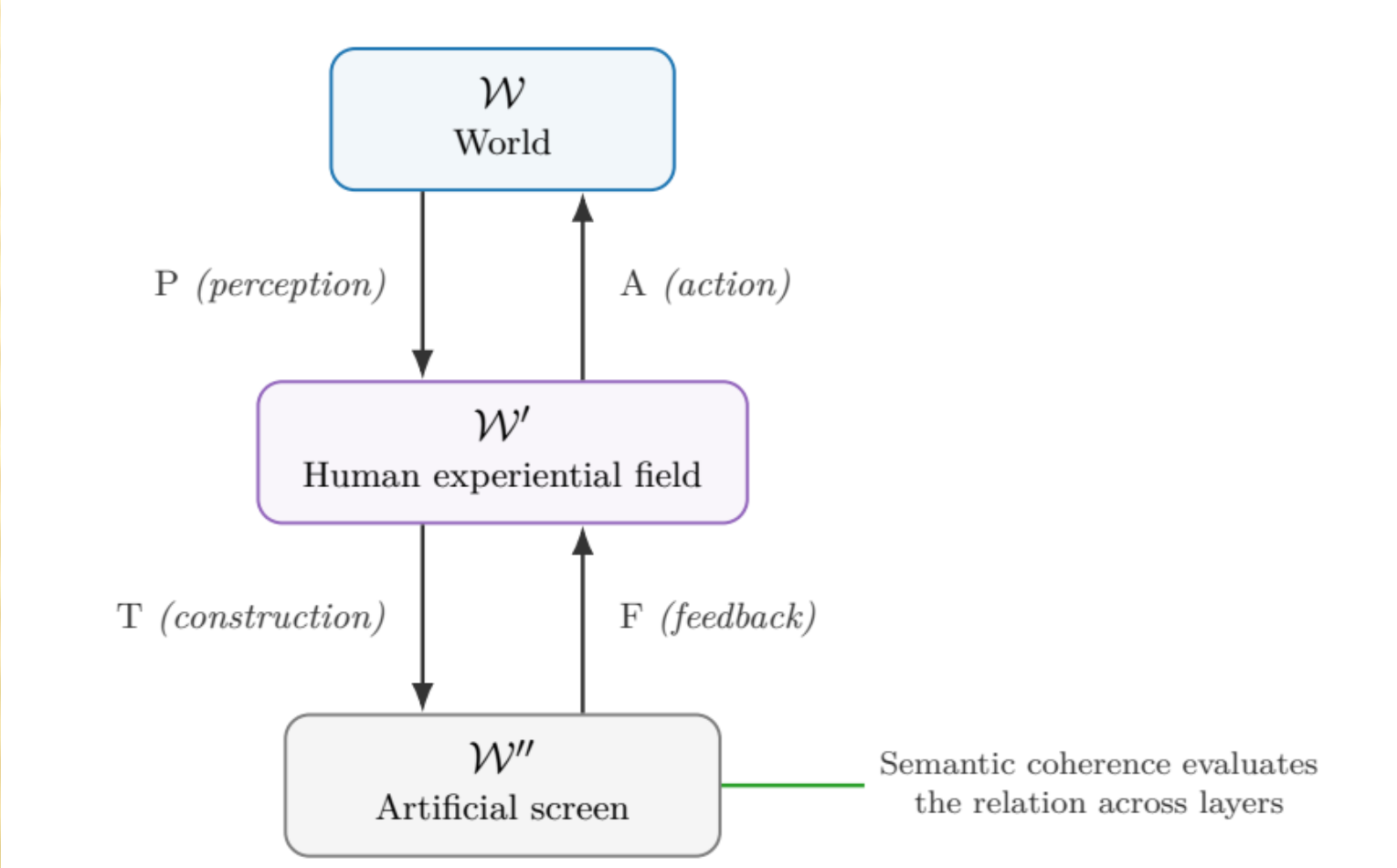
Context-sensitive — The same content can matter differently

Internally grounded — Organization matters for the agent

Meaning can weaken, recover, or recalibrate as representation and agency change over time.

Source figure: Prabhath et al. (2026), Figure 2 and equation 7

Artificial systems add a nested screen



W world

W' human experiential field

W" artificial screen

T technological construction

F feedback into experience

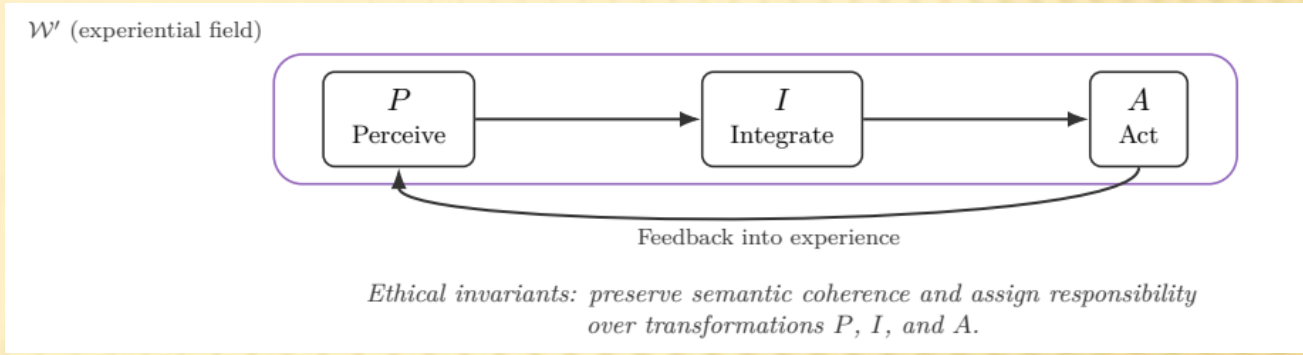
Ethical evaluation concerns the relation across layers, not only the output at W″.

Source figure: Chembrolu et. al. Beyond Alignment, Figure 1

A time-indexed model of mediation

State and design variables

- w_t relevant state of the world
- x_t integrated experiential state
- q_t ranking objective, interface, and personalization policy
- y_t artificial representation constructed under q_t
- F_t uptake of the artificial screen back into experience



$$x_{t+1} = I_t(x_t, P_t(w_t), F_t(y_t))$$

$$w_{t+1} = A_t(x_{t+1}; w_t)$$

A locally permissible y_t changes x_{t+1}, which changes the state used to select the next representation.

Source: Chembrolu et. al. Beyond Alignment, section 3, equations 5–6

Semantic coherence as a diagnostic

Semantic profile

$$\Sigma(z) = (c(z), p(z), s(z), \Omega(z), a(z))$$

- c contextual placement
- p provenance
- s distribution of salience
- Ω reachable interpretations
- a action possibilities available to the agent

Semantic divergence

$$D_{sem}(t) = d_s(\Sigma(x_t), \Sigma(y_t))$$

D_{sem} compares the meaning structure in experience with its technological presentation.

Coherence means an intelligible and corrigible relation between layers. It does not require identity, stasis, or insulation from disagreement.

Source: Chembrolu et. al. Beyond Alignment, section 3, equations 7–8

Perceptual sovereignty – complementary diagnostic

The agent's standing within the process that transforms meaning

$$\pi_t = (\ell_t, \kappa_t, \rho_t, e_t)$$

ℓ	**Legibility**	Can the agent recognize the intervention and its objective?
κ	**Contestability**	Can the agent question the interpretation assigned to them?
ρ	**Redirection**	Can the agent alter parameters or reset the mediation?
e	**Exit**	Can the agent leave without disproportionate cost?

Semantic coherence evaluates the relation between meanings. Perceptual sovereignty evaluates whether the agent can participate in the transformation.

Source: Chembrolu et. al. Beyond Alignment, section 3, perceptual-sovereignty profile

Trajectory-level diagnosis

$$J_{t:t+h} = (\{D_{sem}(\tau)\}, |\Omega(y_{t+h})| - |\Omega(y_t)|, \Delta\pi_{t:t+h})$$

Productive disruption

D_{sem} may rise temporarily

Ω remains stable or expands

π does not deteriorate

Harmful drift

D_{sem} stays elevated or increases

Ω contracts

π declines

The formalism distinguishes coherence from conformity. These are comparative signatures, not validated universal thresholds.

Source: Chembrolu et. al. Beyond Alignment, section 3 and discussion on productive disruption

Two policy-compliant feeds, two trajectories

$$C_t^0 = M_0(C_t)$$

Both feeds draw from the same moderated eligible pool

$y_t^E = Rank(q_t^E; C_t^0, x_t)$

Engagement policy

Shows a cropped clip without date, source, or surrounding exchange

Treats reaction as a preference signal

Makes similar framings increasingly salient

Keeps the inferred profile difficult to inspect or reset

D_{sem} increases, Ω contracts, π declines

$y_t^c = Rank(q_t^c; C_t^0, x_t)$

Coherence-oriented policy

Shows the same clip with source, date, and the full exchange

Keeps alternative interpretations reachable

Explains why the item appears

Lets the user switch objectives or reset personalization

D_{sem} remains corrigible, Ω and π remain available

Both feeds can pass L_0 content moderation at every step.
The engagement trajectory can still fail at L_2.

Source: Chembrolu et. al. Beyond Alignment, section 4, diagnostic recommender example

Four levels of ethical depth

L3	**Foundational**	Conditions for coherence, continuity, perspective, meaning and responsibility
L2	**Experiential**	Semantic coherence, perceptual sovereignty, and representational drift
L1	**Normative**	Rights, values, and intersubjective norms
L0	**Operational**	Behavioral alignment, compliance, documentation, and output metrics

The contribution adds L_2 and L_3 questions without displacing L_0 and L_1 safeguards.

Source: Chembrolu et. al. Beyond Alignment, section 5 and Table 1

The familiar ethical pillars as coherence conditions

Ethical pillar	Representational interpretation
Fairness	Coherence across perspectives and how groups are rendered meaningful
Transparency	Legibility of the transformation and the objective that shaped it
Privacy	Integrity of experiential boundaries and inferred profiles
Accountability	Responsibility for design choices that transform meaning over time
Safety	Preservation of interpretive agency and shared understanding

These interpretations extend operational definitions. They do not replace them.

Source: Chembrolu et. al. Beyond Alignment, section 6

Governance Model Proposal

Before deployment

Map data sources, objectives, interface choices, personalization, memory, and feedback.

During operation

Repeat the assessment with affected groups and compare alternative configurations over time.

01 Contextual fidelity

Are provenance, audience, time, and uncertainty preserved?

02 Semantic legibility

Can users understand why this representation appears?

03 Plurality

Do materially different interpretations remain reachable?

04 Agency and correctability

Can users contest, revise, reset, or exit?

05 Temporal integrity

Does feedback narrow attention or deliberation over time?

No single universal score for meaning: evidence remains application-specific and group-sensitive.

Source: Chembrolu et. al. Beyond Alignment, section 7, preliminary audit scheme

Meaning before alignment

Alignment remains necessary.

Ethical evaluation must also ask whether AI-mediated representations preserve the conditions through which people make sense of the world and act within it.

Acknowledgements

IKSMHA Center, IIT Mandi

Dr. Jyotiranjan Beuria
Vasudeva Prabhath Lolugu
Nitin Kumar

Questions

venkatesh@iitmandi.ac.in

AI & Ethics paper: 10.1007/s43681-026-01355-1