Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Formalising Social Signaling in AI World Models: Structural Limits and Architectural Extensions

Andry Rakotonirainy

r.andry@qut.edu.au

Queensland University of Technology

Mohammed Elhenawy

Queensland University of Technology

Research Article

Keywords:

Posted Date: March 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8896861/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Formalising Social Signaling in AI World Models: Structural Limits and Architectural Extensions

Andry Rakotonirainy (corresponding author)
Mohammed Elhenawy
Queensland University of Technology
{r.andry,mohammed.elhenawy}@qut.edu.au

January 2026

Abstract

Contemporary AI systems increasingly emulate social behaviour, yet their underlying architectures remain primarily epistemic: they optimise for accurate modelling of environmental regularities without incorporating persistent internal goal dynamics that shape agent identity over time. This paper develops a formal account of social signalling within world-model based architectures and identifies structural conditions under which socially grounded meaning can emerge. We show that purely predictive objective functions are insufficient to preserve distinctions that are irrelevant to environmental prediction but essential for social coordination, a limitation called predictive collapse. We then define an architectural extension in which system-relative goal dynamics are coupled to latent state evolution, enabling the stabilisation of socially meaningful representations. A proof-of-concept implementation in a social gridworld illustrates how coupling goals to internal dynamics induces representational differentiation in otherwise prediction-equivalent states. The proposed framework clarifies the computational requirements for integrating social signalling into world models and provides architectural design principles for cognitively grounded artificial social agents.

1 Introduction

Human affinity for a robot tends to increase with human likeness (anthropomorphism), up to a threshold beyond which minor imperfections in social interaction elicit aversion. Such uncanny responses often arise from a mismatch between a robot's outward, observable social behaviour (e.g. a smile) and the absence of corresponding latent and internal affective states [1]. Social signalling, such as a smiling robot, is detached from its necessary human affective ground. It is a technological substitute for socially constituted behaviour. Drawing on Saussurean semiotics, we show that AI systems with embodied features can

generate the outward form of social signals but lack the affective, phenomenal, and socially embedded conditions that constitute genuine social interaction [2]. Current AI architectures treat social signals as outputs to be generated rather than relations constituted within a distributed shared social system. They treat social interaction as a script to be followed rather than a dynamic relationship built between participants.

1.1 Saussure linguistic framework

1.1.1 Definition of Signifier and Signified in social contexts

Saussure stated that linguistic signs consist of two inseparable components: the *signifier* the material form of the sign and the *signified* the concept or meaning to which it refers. Saussure views the signifier sensorial impression associated with a linguistic expression as opposed to a string of letters or an abstract symbol. The signifier is captured from embodied sensory experience such as the sound of spoken language or the visual pattern of written text. It is these experiences that make the signifier cognitively real. The signified is the concept or idea that the signifier points to. For example, the sign *tree*, is the signifier sound pattern or written form "tree", while the signified is the abstract concept of a tree. The relation between the two is arbitrary and socially stabilised: nothing intrinsic to the signifier is fundamentally associated with the signified. The signified is neither a physical object nor a particular mental image. It is a conceptual distinction maintained within a system of differences. Therefore, signification is not reducible to computable perception or statistical prediction and depends on the internal organisation of a social system of signs.

1.1.2 Limitations of AI to express Saussure's social signalling

In Saussure's view, a social signal only has meaning because a group of people have agreed on what it represents and what it entails. Contemporary machine learning (ML) systems, including world models do not instantiate internal, stake-bearing structures that would allow the system itself to participate in a sign relation as defined by Saussure.

In ML systems, a particular action/sign would be generated as a direct optimisation against externally defined objectives (loss, reward, or prompts), making a social function entirely derivative of training criteria. For example, LLMs operate on formal patterns of signifiers without access to the interactive, experiential grounding that connects signifier to signified in human social communication [2]. Therefore, such models manipulate symbols in high-dimensional spaces but they lack the embodied and social context that gives symbols their meaning. In symbolic interactionism, symbols are never self sufficient. They rather acquire significance through lived, embodied exchanges within a community [3].

World models such as JEPA (Joint Embedding Predictive Architecture) [4] go further by learning predictive latent representations of the environment, including social patterns, but these representations remain epistemic rather than

normative. This epistemic neutrality prevents the world model from prioritising social cues. JEPA model may predict when social signals and how they affect others but lacks internal variables for which social signals has intrinsic consequences. Predictive world models hypothesise that machines can learn structured representations of physical reality through self-supervised observation, capturing regularities in sensory data that support accurate prediction of future states. This assumption is related to the statistical structure of the environment as reflected in the data, rather than to the encoding of system-constitutive [1] goals or values. As a result, the "world" represented by such models is epistemic in nature: it specifies what tends to occur, not "what matters" for the system itself. Large language models (LLMs) and world models operate over formal regularities in symbol sequences and sensorimotor data, lacking the intentional and social structures required for semantic content. [5].

Table 1 summarises that existing AI frameworks remains detached from the normative realities of social life. Standard Reinforcement Learning (RL) and LLMs lack a persistent internal state, reducing social interaction to external reactive rewards or statistical mimicry. JEPA (base) has a persistent world model, yet it treats social signals as neutral physical events to be predicted rather than internal values to be upheld. Our extension bridges this "meaning gap" by constitutively coupling the system's internal goals to the social environmental signalling. By encoding social rules/signalling not as external data but as internal constraints essential to the system's identity, this architecture enables actions to carry the weight of shared social reality such as obligation, trust and other social norms.

Epistemic goals (JEPA base) lead to predictive collapse because it optimises away social nuances that don't have immediate physical utility. Only constitutive goals provide the "gravity" needed to keep those social meanings stable.

Architecture	Persistent State	Goal Dynamics	Social Signaling
Standard RL	Limited	External reward	No
LLMs	No	None	Mimicry only
JEPA (base)	Yes	None	Predictive
Our Extension	Yes	Constitutive	Yes

Table 1: Comparison of architectural requirements for socially grounded signification across contemporary AI frameworks.

1.2 World Models to support social signaling

The proposed account of socially grounded signification requires persistent internal state, explicit representation of social context, and system-constitutive goal

[1] System-constitutive denotes goal dynamics that are intrinsic to the identity of the system, persist across contexts, and causally shape internal state evolution, rather than serving as externally defined or task-specific evaluation criteria.

dynamics which are causally dynamic during deployment (as opposed to during training) . Contemporary LLMs mimic social fluency, their underlying architecture is optimised for next-token prediction using externally specified objectives and does not natively support such properties. They operate over transient input contexts and produce outputs via feed forward inference conditioned on fixed parameters. Concepts such as goal, value, or meaning are encoded implicitly through training statistics and is not maintained as an internal variable whose satisfaction or violation has intrinsic consequences for the system. As a result, socially legible behaviours generated by such models remain descriptive or imitative, rather than grounded in internal dynamics that would stabilise social meaning across interactions.

JEPA-like world models are explicitly designed to maintain a latent state that evolves over time. They support anticipation, planning, and reasoning by modelling how latent states of the world evolve over time as a function of actions. Motivation, control and semantic are not treated as intrinsic properties of the predictive representation itself. Therefore world models are more suitable for introducing additional concepts, such as social context and socially stabilised meaning, and for coupling these variables to persistent goal dynamics compared to traditional MLs.

JEPA learn abstract latent representations that capture the regularities and invariances of the sensed environment without relying on explicit generative reconstruction. JEPA's world model is explicitly epistemic. We use the term *epistemic* to denote objectives and internal variables concerned with the accurate representation and prediction of environmental regularities, as opposed to system-constitutive goals or value dynamics that render outcomes intrinsically better or worse for the system itself.

The world model optimisation objective functions are calculated as a learning signal only. After training, it does not correspond to an internal goal whose satisfaction or violation bears consequences for the system itself. As a result, while JEPA-style world models can represent how actions and signals affect future states of the environment, they do not encode which of those futures are preferable or detrimental for the system in a system-constitutive sense. Specifically JEPA's persistent state currently serves as an Internal Physics Engine grounded in universal sensory invariants (e.g. gravitation). It cannot support constitutive concepts for social interaction because these concepts are intersubjective agreements (co-designed) rather than physical laws. While V-JEPA [6] can infer latent information such as gravity from pixel trajectories, it lacks the multimodal feedback loops (deontic/linguistic) required to ground latent social roles, which are negotiated and arbitrary rather than deterministic."

It is important to note that, extending world models to support socially grounded signification is not a trivial modification. It requires abandoning representational neutrality and introducing constitutive coupling between the meaning of variables, contexts, and internal dynamics. These extensions remain conceptually coherent within the world model paradigm, whereas achieving comparable structure within language model architectures would require the introduction of persistent state, internal dynamics, and intrinsic goals that fall

outside their foundational design assumptions.

Our proposed architecture extension is build upon JEPA framework, predictive world models based on self-supervised observation [7]. We make explicit the implications of this architectural choice for social signification. In particular, our extension solves the lack of system-constitutive goal dynamics required to stabilise sign relations in the Saussurean sense. With our extension, the goal and the state are interdependent and transforms the world model from a passive simulator into a normative participant.

This extension follows not from JEPA's insufficient representational capacity or predictive accuracy, but from the absence of intrinsic goal structures within the world model itself. It is therefore important to distinguish between *(i)* the capacity of a system to model the structure of the world and anticipate the consequences of actions, and *(ii)* the capacity of a system to participate in socially grounded sign systems. We argue that the former is directly addressed by JEPA-style architectures; the latter requires additional architectural concepts beyond predictive world modelling.

While motivational or control signals may influence action selection downstream of a world model, they do not, by default, enter into the representational learning objective in a manner that renders goal dynamics constitutive of the latent structure itself. As a result, the predictive representations learned by JEPA-style architectures remain semantically goal-agnostic , even when embedded within a broader agent architecture that includes intrinsic motivation or planning components. This distinction is central to our extension. Grounding social signification requires that representational distinctions be stabilised by their differential consequences for system-constitutive goals. Achieving this would require not merely the addition of motivational modules, but a reorganisation of representational learning such that goal dynamics are intrinsic to, rather than merely consumers of, the learned representations. Our extension does not deny the extensibility of predictive world models, but highlights a structural gap between epistemic world modelling and the requirements of socially grounded signification.

Illustrating example: Robot smiling as an imitation rather than signified signal

An AI system may produce a smile at socially appropriate moments because smiling is statistically associated with positive interaction contexts in its training data. In such cases, the smile reproduces the outward form of a social signal but has no intrinsic consequences for the system itself: whether the smile is perceived as sincere, deceptive, or inappropriate does not alter any persistent internal state. A JEPA-style world model can maintain a latent representation that predicts how smiling affects future behaviour, yet the smile remains epistemic rather than constitutive. It is treated as a regularity to be predicted, not as a signal whose success or failure matters internally. For example, a human smile is meaningful because it carries social risks, such as building or breaking trust. This highlights Saussure's point: mimicking a social being isn't the same

as actually participating in a social system, where your actions have lasting internal consequences.

1.3 Related Work

In structural linguistics, Saussure's original formulation defines the sign as a differential relation within a system of conventions, but does not provide a dynamic or operational account of how signification could be grounded in agent-level processes [2]. Subsequent structural formalisations, such as Hjelmslev's distinction between expression and content planes, further refine the internal structure of signs but remain static and non-dynamic [8]. In semiotics and philosophy of language, meaning has been formalised using logical, code-based, or interpretative frameworks, often treating signification as a mapping between expressions and content governed by shared conventions [9].

More recently, signalling and coordination have been formalised using game theory and information-theory models, in which meaning emerges as an equilibrium property of communicative interactions governed by externally specified utility functions [10, 11]. While these models introduce incentives and optimisation, they treat goals as external and do not account for meaning as grounded in system-constitutive internal dynamics.

Game theory, such as the Nash equilibrium, presupposes shared conceptual structures [12]; equilibrium behavior depends not only on predictive accuracy but on agents' shared recognition of strategies as meaningful and stable [13]. Equilibrium reasoning requires that strategies be identifiable as the same strategy across agents and across time, which already depends on stable conceptual distinctions rather than raw behavioural regularities. Nash equilibrium relies on common knowledge: each agent must not only anticipate others' actions, but recognise that those actions are understood in the same way by all participants. Predictive success alone is not sufficient to secure equilibrium, since an agent may accurately forecast another's behaviour without grasping whether that behaviour counts as cooperative, deceptive, or norm-following. These requirements highlight that coordinated social behaviour depends on shared concepts, not merely on statistical prediction or externally specified objectives.

Our approach to base our concept on system-constitutive goals share similarities with intrinsic motivation in Reinforcement Learning (RL) [14, 15], but differ in that intrinsic rewards such as empowerment remain task-relative as opposed to goals that define system identity. While frameworks such as Homeostatic Reinforcement Learning [16] have successfully internalised physiological stability as a drive for representation learning, they remain limited to the regulation of biological or functional variables. Our work extends this logic into the semiotic domain, arguing that social signals are grounded not in metabolic necessity, but in constitutive social homeostasis, where the 'stability' to be maintained is the agent's standing within a shared system of signification. The symbol grounding problem identifies that symbolic systems require connection to perceptual experience [17]. We extend this with social meaning requiring constitutive coupling to internal goal dynamics – moving beyond the mere association of icons to pix-

els by grounding social signs in the system's own constitutive goals. Work on artificial social intelligence [18] emphasises embodied interaction but typically treats social signals as behavioral outputs rather than meaning-bearing relations grounded in constitutive goals.

Current AI architectures, optimise externally pre-defined objectives without internally grounded concepts and do not directly support shared social understanding mechanisms such as signifier/signified. Note that our approach is not about using Nash equilibrium as a computational target, but as a theoretical lens to illustrate the conceptual prerequisites of coordinated social behavior.

To our knowledge, no existing AI explicitly explicitly integrates Saussure's concepts of sign structure with predictive world models and system-constitutive goal dynamics. The present work aims to bridge this gap by providing a formal account of social signification that is compatible with contemporary world model architectures while remaining faithful to the semiotic structural insights of Saussure.

2 Formalisation

This formulation is constructive and demonstrates feasibility through a minimal proof and implementation section 3

The analysis is grounded in a structural interpretation of signification following Saussure and is articulated in terms of internal state dynamics and goal structures rather than phenomenological assumptions.

We first formalise social signification in a manner faithful to Saussure's structural account while making explicit the role of social context and internal dynamics. We then show how this formalism can, in principle, be integrated into predictive JEPA-type world model architectures. The purpose is not to propose an immediately implementable system, but to identify the additional conceptual structure required. We emphasise that this formulation is constructive rather than prescriptive: it specifies what would be necessary without claiming such structure currently exists in deployed systems.

2.1 Signifier, Signified, and Social Context

Let:

- $\sigma_t \in \Sigma$ denote an observable signal or expression at time t (the *signifier*). It denotes a socially legible expression, such as a smile or linguistic act, which may be derived from x_t through perceptual or behavioural abstraction (e.g., detecting a facial expression from visual input), but is not identified with x_t itself.

- $x_t \in \mathcal{X}$ denote raw sensory observations at time t; such as visual, auditory, or proprioceptive input, prior to any specific task or meaning abstraction. The space $\mathcal{X}$ may be high-dimensional (e.g. pixel arrays or multimodal sensory streams) and is treated as pre-semantic: it contains information

from which social signals may be inferred, but does not itself constitute a signifier in the Saussurean sense. Distinguishing σ_t from x_t allows the same underlying sensory observation to support different social interpretations depending on context, and reflects the fact that signifiers are socially stabilised forms rather than raw sensory data.

- $a_t \in \mathcal{A}$ denote physical actions, decomposable as $a_t = (\sigma_t, a_t^{\text{phys}})$ where σ_t is the socially legible component;
- $z_t \in \mathbb{R}^d$ denote the system's latent representation of the environment and interacting agents;
- $c_t \in \mathcal{C}$ denote a social context variable, encoding socially relevant conditions such as perceived agency, role, interaction history, and normative expectations;
- $\mu_t \in \mathcal{M}$ denote an internal social meaning variable (the *signified*).

The signified is not identified with the full world state, nor with a referential description of the environment. Instead, it represents a socially stabilised abstraction relevant to the interaction. We define the signification relation as a contextualised mapping:

$$\mu_t = h(z_t, \sigma_t, c_t) \tag{1}$$

where h is a contextual abstraction that is many-to-one with respect to physical variation. Multiple physically distinct states may map to the same meaning when they manifest the same socially legible signal under a fixed context, whereas identical signals may yield different meanings when the social context differs:

$$h(z, \sigma, c_1) \neq h(z, \sigma, c_2) \quad \text{in general.}$$

This context dependence is essential: signification is not a property of the signal alone, but of the signal as embedded within a socially instantiated system of differences.

2.2 Standard JEPA World Model

Having established the formal requirements for social signification, we now examine how contemporary JEPA-style world models relate to these requirements and what extensions would be necessary to support socially grounded signalling.

In JEPA, observations $x_t \in \mathcal{X}$ at time t are encoded to latent states $z_t \in \mathbb{R}^d$ via an encoder $f_\theta : \mathcal{X} \to \mathbb{R}^d$:

$$z_t = f_\theta(x_t)$$

Future latent states are predicted as a function of the current latent state and action $a_t \in \mathcal{A}$ via a predictor network $g_\phi : \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}^d$:

$$\hat{z}_{t+1} = g_\phi(z_t, a_t)$$

The model is trained using a purely *epistemic* (self-supervised) objective that minimises prediction error in latent space, without requiring task-specific rewards:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}\left[d\left(\mathrm{sg}(f_{\bar{\theta}}(x_{t+1})),\, g_\phi(f_\theta(x_t), a_t)\right)\right] \tag{2}$$

where $d(\cdot, \cdot)$ is a distance metric (e.g., mean squared error), $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator that blocks backpropagation, and $\bar{\theta}$ is an exponential moving average (EMA) of the online encoder parameters: $\bar{\theta} \leftarrow \mu\bar{\theta} + (1-\mu)\theta$ with momentum $\mu \approx 0.99$.

This asymmetric architecture prevents representational collapse by ensuring the target encoder $f_{\bar{\theta}}$ is not directly optimised by the prediction loss. The epistemic objective enables learning world dynamics independently of downstream task rewards.

While JEPA learns how the world evolves, including social regularities, it lacks internal variables for which social signals have intrinsic consequences. The loss $\mathcal{L}$ functions only as a training signal and does not correspond to persistent, system-relative goals that remain causally active during deployment.

2.3 Extension: Incorporating Context and Meaning

To support Saussurean social signalling, the world model must be extended to represent social context explicitly. We introduce a social context variable $c_t \in C$, which encodes socially relevant conditions such as perceived roles, interaction history, and normative expectations. While social context may depend on past observations and actions, the context encoder does not require access to the full sensory or action history. Instead, c_t is derived from the current observation together with a maintained internal summary H_t of prior interactions:

$$c_t = f_\psi(x_t, H_t) \tag{3}$$

where H_t aggregates socially relevant aspects of past experience and is updated incrementally over time.

The predictive world model then represents joint transitions in world state and social context as

$$P(z_{t+1}, c_{t+1} \mid z_t, c_t, a_t) \tag{4}$$

Here, the current action a_t is included explicitly because it constitutes a causal intervention that drives state evolution at time t, rather than merely a historical factor. Past actions influence future behaviour indirectly through their contribution to the maintained context summary H_t and hence to c_t, whereas a_t determines the immediate transition of the system and its environment. This distinction separates the role of action in shaping social context over time from its causal role in driving moment-to-moment dynamics, and will be central to the discussion in Section 3 when interpreting the architectural implications of constitutive goal coupling.

Unlike z_t, which is optimised solely for predictive accuracy via $\mathcal{L}(\theta, \phi)$ the meaning variable μ_t must be shaped by its consequences for system-constitutive

goals. This requires coupling meaning to goal dynamics, as specified below, and cannot be achieved through epistemic prediction alone.

2.4 Extension: System-Constitutive Goal Dynamics

Standard JEPA dynamics (Eq. 4) are purely epistemic: they predict what will occur, not what matters for the system. To support social signification, we propose that internal state evolution must depend on goal-relevant meaning.

By *system-constitutive goal dynamics* we mean goal-oriented structures that: (*i*)are intrinsic to the system (not externally imposed); (*ii*) persist across contexts and tasks; (*iii*) continue to dynamically modulate internal state evolution during deployment; (*iv*) cannot be removed without altering the system's functional identity. For example, G might represent trust, social affiliation, or reputation—internal variables whose satisfaction or violation has enduring consequences for future interactions.

Formally, a system exhibits system-constitutive goal dynamics if there exists a goal functional:

$$G : \mathcal{M} \times \mathcal{C} \to \mathbb{R}^k$$

such that variations in $G(\mu_t, c_t)$ represent outcomes that are better or worse for the system itself, and G remains causally active in determining state evolution:

$$(z_{t+1}, c_{t+1}) = g^{\text{const}}(z_t, c_t, a_t, G(\mu_t, c_t)) \tag{5}$$

where g^{const} denotes the constitutive dynamics function (distinguished from the epistemic predictor g_ϕ in Eq.2).

G is not optimised solely during training, nor does it function merely as an evaluative signal for action selection. It constrains the system's internal dynamics across time, ensuring that counterfactual variations in signalling lead to systematically different internal states/trajectories.

3 Validation

We validate the proposed extension through a combination of theoretical analysis, demonstration and prototyping. Specifically, we (i) establish a formal limitation of purely epistemic world models, showing that standard JEPA-style objectives cannot guarantee the representation of goal-relevant social distinctions, (ii) clarify how the proposed formalisation introduces concepts that go beyond epistemic prediction, and (iii) provide a minimal prototype illustrating the representational consequences of coupling internal goal dynamics to a JEPA architecture.

3.1 Distinguishing Constitutive Dynamics from Auxiliary Tasks

We make a critical distinction between system-constitutive dynamics and the use of auxiliary tasks in standard multi-task learning. One might argue that

social grounding could be achieved by adding an auxiliary loss function L_{aux} to a standard JEPA architecture. However, it could result in a fundamentally different representational structure.

In an auxiliary-task configuration, an encoder f_θ is typically optimised as follows:

$$\min_{\theta,\phi} \mathbb{E}[L_{pred}(z_{t+1}, \hat{z}_{t+1}) + \lambda L_{aux}(z_t, y_{social})]$$

where y_{social} represents an externally provided social label or reward. In this setup, the social variable is merely a *consumer* of the representation; it provides a gradient during training but does not fundamentally alter the transition logic of the world model during deployment. The resulting latent space remains **epistemic**, as it is shaped to accurately predict external labels rather than internalise the normative significance of the signal. Here, the influence of an auxiliary objective is confined to the training phase. During deployment, L_{aux} is no longer evaluated, and the system does not internally represent whether social agreements/obligations are satisfied or violated. Any normative regularities learned via auxiliary supervision persist only implicitly in the fixed parameters and exert no direct causal influence on state transitions at runtime. In this sense, auxiliary objectives bias representations offline but do not constitute goals that are monitored, maintained, or enforced by the system itself.

In contrast, our constitutive framework (Eq. 5) embeds the goal functional G directly into the transition dynamics:

$$(z_{t+1}, c_{t+1}) = g^{const}(z_t, c_t, a_t, G(\mu_t, c_t))$$

where g^{const} denotes the constitutive dynamics function. Here, the goal is not a static evaluation metric but a **state modulator** that is causally active during deployment. The structural implication is twofold:

- Causal Integration: The system's internal state evolution is physically inseparable from its social goals. If the system fails to maintain goal homeostasis (e.g., trust or affiliation), its future representational trajectories are systematically altered.
- Normative Grounding: Unlike an auxiliary task which relies on external labels, constitutive goals define the system's functional identity. This moves the system from *mimicry* (predicting what social signals occur) to *signification* (internalising what those signals mean for the system's own persistence).

3.2 Representational Limitation of Epistemic Objectives

A fundamental consequence of optimising purely epistemic objectives is that representations are shaped solely by predictive utility. If two internal states z_A, z_B lead to identical predicted futures, they will be mapped to equivalent representations, even if they differ in goal-relevant meaning. This *predictive collapse* explains why JEPA models can accurately forecast social dynamics while remaining semantically neutral regarding their normative significance.

Proposition Let $L : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ be a prediction loss, and let $f_\theta : \mathcal{X} \to \mathcal{Z}$ be an encoder optimised to minimise $\mathbb{E}[L(z_{t+1}, \hat{z}_{t+1})]$. If two latent states $z_A, z_B \in \mathcal{Z}$ satisfy

$$L(z_{t+1}, g(z_A, a)) = L(z_{t+1}, g(z_B, a)) \quad \text{for all } a \in \mathcal{A}$$

then gradient descent on L provides no signal to distinguish z_A from z_B, regardless of their goal-relevant semantics.

Proof. Since $\partial L/\partial\theta$ depends only on the prediction error $L(z_{t+1}, \hat{z}_{t+1})$, representations that yield identical predictions produce identical gradients. Thus, goal-relevant distinctions orthogonal to predictive accuracy are not learnable under purely epistemic objectives. □

Concretely, social distinctions such as the difference between a "sincere" smile and a "deceptive" one become unlearnable if they do not manifest as divergent physical trajectories in the short term.

3.3 Illustrative Example: Trust Signaling in a Social Dilemma

To illustrate the distinction between epistemic prediction and system-constitutive grounding, consider a simplified trust game between two agents A and B. Agent A can send a signal $\sigma_t \in \{\text{cooperate}, \text{defect}\}$ to B. A standard JEPA world model can learn:

$$P(z_{t+1}|z_t, \sigma_t)$$

predicting that "cooperate" signals often precede cooperative responses from B. This representation is purely epistemic. The model encodes *what B typically does* when receiving "cooperate", but has no internal variable for which the response matters *to A itself*.

The structural limitation manifests in several ways:

- Deceptive alignment: Models may learn to produce socially appropriate signals (politeness, agreement, smiles) that predict positive human feedback, without these signals connecting to internal states that would make deception costly to the system.
- Reward hacking: Reinforcement Learning agents optimise externally specified rewards but lack intrinsic goals that would prevent gaming the metric. This demonstrates that task-specified objectives, even when sophisticated, do not create system-constitutive stakes.

These phenomena support the claim that predictive accuracy, even over social interactions, is insufficient for socially grounded signification as per Saussure's theory. This demonstration shows what empirical signature would distinguish epistemic from constitutive architectures.

3.4 Prototyping

This section provides a prototype of the formal specification developed in previous sections. It is no an empirical refutation of all possible JEPA instantiations. The theoretical results in the validations Section 3.1 establish that,

under purely epistemic objectives, JEPA-style world models cannot systematically encode distinctions whose relevance is orthogonal to predictive accuracy. Here, we demonstrate how introducing system-constitutive goal dynamics yields representational effects that are unavailable to standard epistemic JEPA under identical predictive conditions.

Using the official Meta I-JEPA implementation as an epistemic baseline [19]. We modified the architecture only to introduce a persistent, causally active internal goal variable, while leaving the predictive objective unchanged. This controlled comparison is designed to isolate the structural consequences of constitutive coupling, rather than to optimise task performance or improve prediction accuracy. The purpose of this prototype is therefore to show, by construction, that embedding goal dynamics within internal state evolution produces stable representational differentiation in social contexts. This simulate exactly the form of differentiation that purely epistemic world models are formally unable to guarantee.

The resulting implementation should be understood as a pseudo proof for the proposed extension, confirming that the formal specification described in Section 2 are both coherent and practically realisable within the JEPA framework.

Figure 1 illustrates the pipeline architecture associated with the formal specification. The architecture highlights that socially grounded signification requires an explicit separation between observable signals, contextualised meaning variables, and system-constitutive goal dynamics.

Environment: We implemented a minimal social gridworld with 5×5 grid with ego agent and social partner. Partner sends signals (approach/avoid) and performs actions (help/harm). Signal-action correlations vary by context.

Architectures:

- Epistemic JEPA: Standard encoder-predictor trained to minimise $\mathcal{L}(z_{t+1}, \hat{z}_{t+1})$
- Constitutive JEPA: Standard JEPA architecture + persistent trust variable τ_t coupled to encoder: $z_t = f_\theta(x_t, \tau_t)$

Metrics:

- Representational differentiation: $|z_{\text{high trust}} - z_{\text{low trust}}|$ for identical signals
- Adaptation score: Deviation of τ_t from neutral after interactions

Both models used 4-dimensional latent spaces, linear encoders, and EMA target encoders were used with decay parameter $\mu = 0.99$. In the constitutive variant, a scalar trust variable $\tau \in [0, 1]$ was coupled to the encoder via multiplicative gating.

$$z = f_\theta(x)\left(0.5 + 0.5\,\tau\right)$$

Trust was updated as a function of interaction outcomes, with

$$\Delta\tau = \begin{cases} +0.1, & \text{for helpful outcomes,} \\ -0.2, & \text{for harmful outcomes.} \end{cases}$$

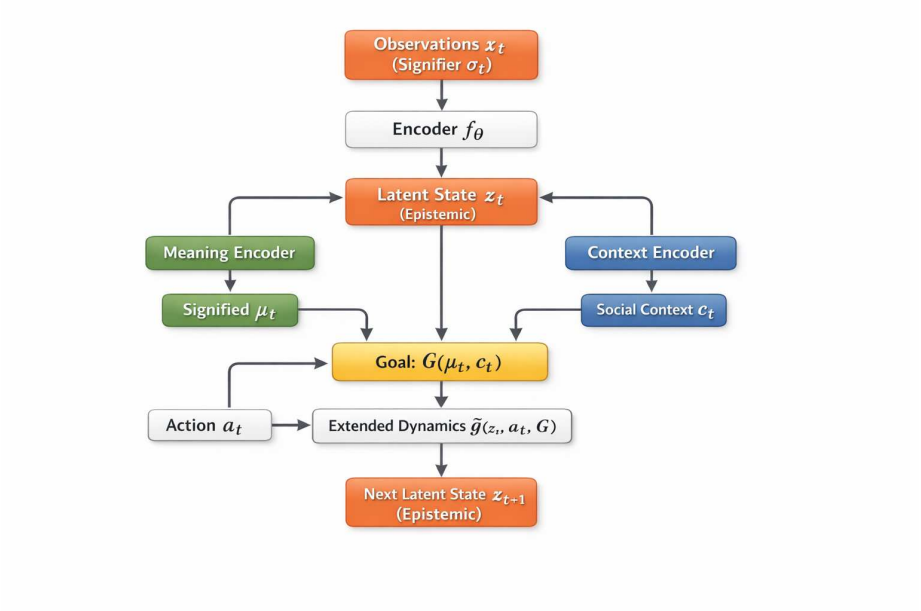


Figure 1: Extension of JEPA world model architecture with contextualised social signification and system-constitutive goal dynamics. Observable signals (signifiers) are distinguished from epistemic world representations, and social meaning is grounded through persistent internal goal dynamics.

The current implementation updates the trust variable τ via discrete increments (Eq. 6), this serves as a minimal proof-of-concept for representational differentiation. Ideally, τ would emerge through the same self-supervised mechanisms that JEPA uses for physical invariants, but grounded in inter-subjective feedback loops (e.g., linguistic validation or social cooperation) rather than prediction alone.

3.4.1 Simulation results

Figure 2 contrasts an epistemic JEPA with a constitutive variant that includes a trust-gated variable. Panel (A) shows that the epistemic JEPA maintains negligible representational differentiation across episodes, consistent with a latent state optimised solely for prediction. In contrast, the constitutive model exhibits a stable, non-zero level of differentiation, indicating that representations are modulated by prior outcomes rather than prediction error alone. Panel (B) shows that the internal trust state rapidly departs from neutrality and stabilises at a lower regime, reflecting outcome-based normative evaluation rather than optimisation for predictive accuracy. Panel (C) confirms that both models converge to comparable prediction loss, ruling out differences in learning efficiency or capacity as an explanation. Panel (D) summarises the resulting separation at

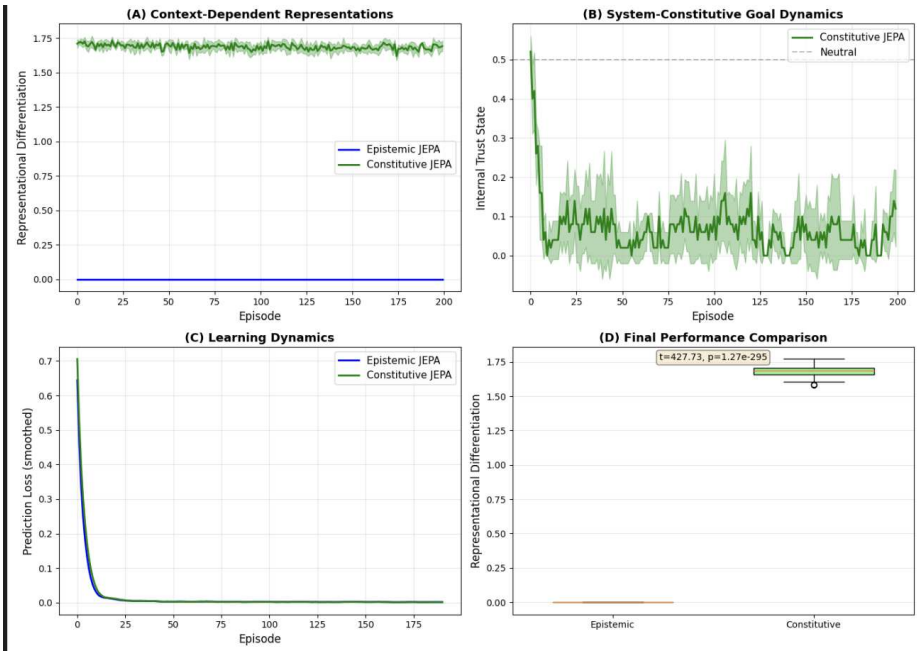


Figure 2: Comparison of epistemic and constitutive JEPA variants. (A) The epistemic JEPA shows negligible representational differentiation across episodes, while the constitutive model maintains a stable, non-zero differentiation. (B) The trust variable departs from neutrality and stabilises based on outcome-driven updates. (C) Both models converge to comparable prediction loss. (D) Despite equivalent predictive performance, the constitutive model exhibits significantly higher final representational differentiation.

convergence: despite equivalent predictive performance, the constitutive model maintains significantly higher representational differentiation. Together, these results indicate that the observed advantage arises not from improved prediction but from a structural change in how signals are taken up, supporting the claim that persistent predictive state alone is insufficient for constitutive social meaning. We have confirmed that coupling goals to dynamics fundamentally changes what representations are learned, supporting the theoretical claim that epistemic prediction is insufficient for socially grounded signification.

These preliminary results support the claim that coupling goals to internal dynamics changes *what* representations are learned. The epistemic baseline model learns signal-outcome correlations but collapses different contexts. It exhibits predictive collapse with a near-total overlap in representations. In contrast, our constitutive model differentiates contexts because they have different implications for internal trust dynamics. Our model forces a stable representational divergence, demonstrating that the architecture grounds social meaning as a structural feature of its world model.

While simplified, this demonstrates that the architectural distinction is empirically meaningful and could guide future implementations.

4 Conclusion

This paper identifies a fundamental oversight in current AI approach: the belief that "better perception" leads to "better social behavior". We have shown that the inability of modern agents to engage in genuine social signalling stems not from a lack of sensory data, but from the structural absence of system-constitutive goal dynamics. While current world models are excellent at predicting how a social partner might behave, they lack any internal mechanisms reflecting what is at stake for the system itself.

By formalising an architectural extension to the JEPA framework, we bridge the gap between purely epistemic prediction and socially grounded meaning (à la Saussure). Our approach represents a substantive departure from traditional AI architectural design. Goals are no longer mere downstream evaluators for planning, but are active participants that shape the system's internal state evolution in real-time . This coupling forces the emergence of context-sensitive meaning structures that go beyond simple statistical mimicry.

This research does not seek to replace the epistemic strengths of existing world models, but to expand them. The path forward requires us to move beyond the comfort of self-supervised learning for accuracy alone. The next major challenge for AI is the development of training objectives that prioritise social homeostasis and long-term goal satisfaction by creating agents that do not just act socially, but participate in the normative reality of a shared social world.

Funding

This research was supported by the Australian Research Council (ARC) under the Discovery Project scheme (DP220102598).

Data Availability

No external datasets were used in this study. The simulation code supporting the findings of this study is available from the corresponding author upon reasonable request.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

A.R. conceived and designed the study, developed the formal framework, and drafted the manuscript. M.E. contributed substantially to the formalisation of the work and critically revised the manuscript for important intellectual content. Both authors approved the final version of the manuscript.

References

[1] Mori M, MacDorman KF, Kageki N. The Uncanny Valley [From the Field]. IEEE Robotics & Automation Magazine. 2012;19(2):98-100.

[2] de Saussure F. Cours de linguistique générale. Bally C, Sechehaye A, editors. Lausanne and Paris: Payot; 1916.

[3] Blumer H. Symbolic Interactionism: Perspective and Method. University of California Press; 1969.

[4] Monemi M, Chinipardaz M, Rasti M, Bennis M, Latva-Aho M. Tutorial on Joint Embedding Predictive Architectures (JEPA): Foundations, Applications, and Future Directions. Preprints; 2025. Available from: https://www.techrxiv.org/users/866579/articles/1365143-tutorial-on-joint-embedding-predictive-architectures-jepa-foundations-applicati commit=cf8057b70c0ce83e6bfbc62c6d620d3e6cf81fdd.

[5] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 610-23. Available from: https://dl.acm.org/doi/10.1145/3442188.3445922.

[6] Assran M, Bardes A, Fan D, Garrido Q, Howes R, Mojtaba, et al.. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. arXiv; 2025. ArXiv:2506.09985 [cs]. Available from: http://arxiv.org/abs/2506.09985.

[7] Lecun Y. A path towards autonomous machine intelligence. Courant Institute of Mathematical Sciences, New York University and Meta – Fundamental AI Research; 2022. Version 0.9.2). Available from: https://openreview.net/pdf?id=BZ5a1r-kVsf.

[8] Hjelmslev L. Prolegomena to a Theory of Language. University of Wisconsin Press; 1961.

[9] Eco U. A Theory of Semiotics. Indiana University Press; 1976.

[10] Lewis D. Convention: A Philosophical Study. Harvard University Press; 1969.

[11] Skyrms B. Signals: Evolution, Learning, and Information. Oxford University Press; 2010.

[12] Aumann RJ. Agreeing to Disagree. The Annals of Statistics. 1976;4(6):1236-9. Available from: https://www.jstor.org/stable/2958591.

[13] Myerson RB. Nash Equilibrium and the History of Economic Theory. Journal of Economic Literature. 1999 Sep;37(3):1067-82. Available from: https://www.aeaweb.org/articles?id=10.1257/jel.37.3.1067.

[14] Schmidhuber J. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). IEEE Transactions on Autonomous Mental Development. 2010;2(3):230-47.

[15] Oudeyer PY, Kaplan F. What is Intrinsic Motivation? A Typology of Computational Approaches. Frontiers in Neurorobotics. 2007 Nov;1:6. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC2533589/.

[16] Keramati M, Gutkin B. Homeostatic reinforcement learning for integrating reward collection and physiological stability. eLife. 2014 Dec;3:e04811.

[17] Harnad S. The symbol grounding problem. Physica D: Nonlinear Phenomena. 1990 Jun;42(1):335-46. Available from: https://www.sciencedirect.com/science/article/pii/0167278990900876.

[18] Dautenhahn K. Socially intelligent robots: dimensions of human–robot interaction. Philosophical Transactions of the Royal Society B: Biological Sciences. 2007 Feb;362(1480):679-704. Available from: https://doi.org/10.1098/rstb.2006.2004.

[19] Assran M, Duval Q, Misra I, Bojanowski P, Vincent P, Rabbat M, et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. p. 15619-29.